Dissertação apresentada à Pró-Reitoria de Pós-Graduação do Instituto Tecnológico de Aeronáutica e da Universidade Federal de São Paulo, como parte dos requisitos para obtenção do título de Mestre em Ciências no Programa de Pós-Graduação em Pesquisa Operacional, Área de Gestão e Apoio a Decisão.

**Rosana Batista Teixeira**

# ANTECIPAÇÃO DE MUDANÇA DE REGIME NA FATIA DIÁRIA DE VOOS ATRASADOS E CANCELADOS NO AEROPORTO INTERNACIONAL DE GUARULHOS

Dissertação aprovada em sua versão final pelos abaixo assinados:

Prof. Dr. Rodrigo Arnaldo Scarpel
Orientador

Prof. Dr. Pedro Teixeira Lacava
Pró-Reitor de Pós-Graduação

Campo Montenegro
São José dos Campos, SP – Brasil
2019

**Dados Internacionais de Catalogação-na-Publicação (CIP)**
**Divisão de Informação e Documentação**

Teixeira, Rosana Batista
ANTECIPAÇÃO DE MUDANÇA DE REGIME NA FATIA DIÁRIA DE VOOS ATRASADOS E CANCELADOS NO AEROPORTO INTERNACIONAL DE GUARULHOS/ Rosana Batista Teixeira.
São José dos Campos, 2019.
97f.

Dissertação de mestrado – Curso de Pesquisa Operacional. Área de Gestão e Apoio a Decisão – Instituto Tecnológico de Aeronáutica e Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo, 2019. Orientador: Prof. Dr. Rodrigo Arnaldo Scarpel.

1. Modelos escondidos de Markov. 2. Classificações. 3. Operações de Linhas aéreas. I. Instituto Tecnológico de Aeronáutica. II. Universidade Federal de São Paulo. III. Título

## REFERÊNCIA BIBLIOGRÁFICA

TEIXEIRA, Rosana Batista. **ANTECIPAÇÃO DE MUDANÇA DE REGIME NA FATIA DIÁRIA DE VOOS ATRASADOS E CANCELADOS NO AEROPORTO INTERNACIONAL DE GUARULHOS**. 2019. 97f. Dissertação de mestrado em Pesquisa Operacional – Instituto Tecnológico de Aeronáutica e Universidade Federal de São Paulo, São José dos Campos.

## CESSÃO DE DIREITOS

NOME DO DA AUTORA : Rosana Batista Teixeira
TÍTULO DO TRABALHO: ANTECIPAÇÃO DE MUDANÇA DE REGIME NA FATIA DIÁRIA DE VOOS ATRASADOS E CANCELADOS NO AEROPORTO INTERNACIONAL DE GUARULHOS.
TIPO DO TRABALHO/ANO: Dissertação / 2019



_____

Rosana Batista Teixeira
Av Gaspar de Souza, Praia das Palmeiras
11.666-250 – Caraguatatuba -SP

# ANTECIPAÇÃO DE MUDANÇA DE REGIME NA FATIA DIÁRIA DE VOOS ATRASADOS E CANCELADOS NO AEROPORTO INTERNACIONAL DE GUARULHOS

**Rosana Batista Teixeira**

Composição da Banca Examinadora:

| | | | | |
|---|---|---|---|---|
| Prof$_a$. Dr$_a$. | Denise Beatriz T. P. A. Ferrari | Presidente | - | ITA |
| Prof. Dr. | Rodrigo A. Scarpel | Orientador | - | ITA |
| Prof. Dr. | Ana Carolina Lorena | | - | ITA |
| Prof. Dr. | Rafael Duarte C. dos Santos | | - | INPE |

**ITA**

À minha mãe e aos meus filhos pela compreensão, incentivo e apoio incondicional.

# Agradecimentos

Em primeiro lugar gostaria de agradecer a Deus por colocar pessoas especiais no meu caminho e pela força em momentos difíceis.

Ao meu orientador, Professor Doutor Rodrigo A. Scarpel, pelo encorajamento enquanto dava meus primeiros passos na pós-graduação. Obrigada pela paciência, confiança e ensinamentos compartilhados na realização deste trabalho. Serei sempre grata!

À minha família por todo suporte. Obrigada por acreditarem e assumirem comigo esta jornada. Agradeço a minha mãe, Ana, que mesmo distante é minha grande companheira de todas as horas. Ao meu padrasto, Aldo, por todo apoio. Aos meus filhos Gabriel e Danna que não hesitaram em me apoiar para que eu pudesse chegar ao fim desta etapa. Amo vocês!

Ao meu namorado Renê pelo carinho, compreensão, contribuição, experiências compartilhadas e especialmente pelo apoio no início desta trajetória.

À minha amiga Cristina e meu cunhado Renato pelo tempo dedicado aos meus filhos nesta fase. Às minhas amigas Emília, que muito contribuiu para que obstáculos pudessem ser superados, e Jaqueline, grande companheira. Ao meu colega Nícolas por me receber sempre que precisei.

Ao ao meu amigo Guilherme por acreditar, estar sempre presente, pela paciência ao compartilhar seu conhecimento e pelo estímulo nos momentos difíceis.

Por fim, agradeço a todos que de alguma forma contribuíram para a realização deste trabalho!

*"Se você estiver enfrentando um novo desafio*
*ou for solicitado a fazer algo que nunca fez antes,*
*não tenha medo. Tem mais capacidade do que pensa,*
*mas nunca a verá, a menos que faça uma exigência a si mesma*
*por mais."*
— Joyce Meyer

# Resumo

Atrasos e cancelamentos de voos são ocorrências frequentes na maioria dos aeroportos em todo o mundo. No Brasil, o aumento desregulamentado do tráfego aéreo provocou a concentração de voos em alguns aeroportos e possibilitou a ocorrência de atrasos e cancelamentos de voos em razão de dias congestionados. Dentre estes aeroportos, o Aeroporto Internacional de Guarulhos (GRU) é o mais afetado por atrasos no país. Portanto, o objetivo deste trabalho é a criação de um modelo de previsão que visa antecipar a ocorrência de dias congestionados no Aeroporto Internacional de Guarulhos. Para a composição do modelo foram empregues os Modelos Escondidos de Markov, como uma abordagem de agrupamento, e três classificadores: Árvore de Classificação e Regressão, Florestas Aleatórias e Máquina de Vetores de Suporte. A precisão do modelo foi considerada satisfatória e antecipou a mudança de regime na fatia diária por um período a frente.

# Abstract

Flight delays and cancellations are frequent occurrences in most airports around the world. In Brazil the deregulated increase in air traffic caused flight concentration in some airports, enabling the occurrence of delays and cancellations due to congested days. The Guarulhos International Airport is the most affected by delays. Therefore, the goal of this work is to anticipate the occurrence of congested days at Guarulhos International Airport employing clustering and classification approaches to identify a regime change in the daily share of delayed and canceled flights. The built model is composed of a Hidden Markov Models as a clustering approach and the classification methods Classification and Regression Tree, Random Forest, and Support Vector Machine. The accuracy of the prediction model was considered satisfactory, and it was able to anticipate the regime change in a daily share for one period ahead.

# Lista de Figuras

# Lista de Tabelas

# Lista de Abreviaturas e Siglas

| | |
|---|---|
| AIC | Akaike Information Criterion |
| ANAC | Agência Nacional de Aviação Civil |
| AUC | Area Under the Curve |
| BIC | Bayesian Information Criterion |
| CART | Classification and Regression Trees |
| CDLL | Complete-data Log-likelihood |
| CPD | Change Point Detection |
| DECEA | Departamento de Controle do Espaço Aéreo |
| DM | Data Mining |
| EM | Expectation-Maximization |
| FAA | Federal Aviation Administration |
| GRU | Aeroporto Internacional de Guarulhos |
| HHI | Herfindal-Hirschman Index |
| HMM | Hidden Markov Model |
| Ibovespa | Bolsa de Valores de São Paulo |
| KDD | Knowledge Discovery in Databases |
| LL | Log-Likelihood |
| MC | Markov Chain |
| MT | Matriz de Transição |
| OOB | Out of Bag |
| RF | Random Forest |
| ROC Curve | Receiver Operating Characteristic Curve |
| SVM | Support Vector Machine |

# Sumário

# 1 INTRODUÇÃO

A ocorrência de dias congestionados nos aeroportos em decorrência de atrasos e cancelamentos de voos é um problema universal. De acordo com Bendinelli *et al.* (2016) atrasos se tornaram uma realidade na indústria aérea global. Atrasos e cancelamentos de voos têm sido objeto de estudos científicos e muitos pesquisadores estão consoantes em afirmar que o desbalanceamento entre demanda e capacidade tem forte relação com os atrasos e cancelamentos de voos. De acordo com Xiong e Hansen (2013), o sistema de aviação enfrenta grandes desafios ao lidar com a alta demanda quando a capacidade do sistema é reduzida. Diante dos atrasos, os horários das companhias aéreas podem sofrer mudanças não previstas, pois alguns voos se atrasam em razão da chegada tardia do voo anterior e, devido aos horários apertados, estes atrasos podem se propagar (ABDEL-ATY *et al.*, 2007).

Ferguson *et al.* (2013) afirmam que existem duas razões como causas principais dos atrasos de voos: a primeira são os voos não partirem devido a ocorrer algum problema na aeronave ou no voo (problemas mecânicos e regras de trabalho da tripulação são alguns exemplos); e a segunda razão é a incompatibilidade entre a demanda e a capacidade dos aeroportos. De acordo com Jacquillat e Odoni (2015), a maioria dos atrasos de voos são resultantes do desbalanceamento entre demanda e capacidade. Segundo os autores, este desequilíbrio é causado pelo crescimento do tráfego aéreo e as limitações de capacidade dos aeroportos com grande movimento.

Não obstante haver consonância ao se tratar do desbalanceamento entre demanda e capacidade, são apresentadas na literatura inúmeras causas, por perspectivas diferentes, para que ocorram incompatibilidades entre a demanda e capacidade dos aeroportos. De acordo com Xiong e Hansen (2013), os atrasos são problemas significantes, resultantes da excessiva demanda de voos e estão fortemente associados com as operações, duração do voo e condições climáticas dos aeroportos de origem destino. Os autores reiteram que este problema é agravado pela competitividade das companhias aéreas, que, confrontadas pelos altos custos das aeronaves, buscam o máximo aproveitamento. Ainda segundo os autores, como estratégia, as companhias aéreas aumentam atrasos de alguns voos, reduzem em outros e cancelam voos para evitarem atrasos muito extensos ou para desocuparem espaço para outros voos.

Santos *et al.* (2018) afirmam que atrasos e cancelamentos de voos apresentam-se como alguns dos principais problemas associados à interrupção das operações de uma rede de transporte aéreo. Janić (2015) afirma que uma rede de transporte aéreo consiste em aeroportos e rotas operadas pelas companhias aéreas. Os aeroportos representam nós, e as rotas, as ligações entre os aeroportos. Segundo o autor, as perturbações de grande escala podem comprometer o funcionamento da rede. Ente elas estão o mau tempo, falhas de determinados componentes da rede consideradas cruciais (sistemas dos computadores centrais), ações relacionadas aos funcionários de transporte aéreo (por exemplo, greves), desastres naturais, ameaças e ataques terroristas, incidentes ou acidentes aéreos.

De acordo com Abdel-Aty *et al.* (2007), o aumento do atraso de voos deve-se principalmente ao clima adverso nas proximidades dos aeroportos, a falta de capacidade das pistas, ao aumento do número de aeronaves e ao controle de tráfego aéreo deficiente. Por uma concepção diferente, a partir de um relatório da *Federal Aviation Administration* (FAA) de 2014, Bendinelli *et al.* (2016) concluíram que a ausência de concorrência favorecia o aumento das taxas de atrasos e cancelamento de voos.

Para tratar atrasos de voos, Jacquillat e Odoni (2015) utilizaram técnicas de programação inteira e dinâmica para o desenvolvimento e aplicação de uma abordagem que otimiza as intervenções nos agendamentos e a utilização da capacidade aeroportuária, ou seja, como os procedimento operacionais podem ser modificados para minimizar os custos de congestionamento em nível tático. Os autores desta abordagem. Madas e Zografos (2008) desenvolveram uma estrutura na qual utilizaram estratégias alternativas de alocação de slots em diferentes configurações de aeroportos europeus para tratar a escassez de capacidade nos aeroportos, da qual são resultantes congestionamentos e atrasos de voos.

Santos e Robin (2010) utilizaram análise de regressão múltipla para identificar as causas dos atrasos nos aeroportos europeus. Para o Aeroporto Internacional de Guarulhos, Scarpel e Pelicioni (2018) desenvolveram um modelo de alerta de antecipação de atrasos baseado na combinação de indicadores de alerta contra a ocorrência de mudanças em uma variável de interesse (fatia diária de movimentos totais, ou seja, chegadas e partidas ocorridos com atraso). Foram considerados atrasos, de acordo com as normas internacionais, os voos que chegassem ou partissem com mais de quinze minutos do horário previsto.

## 1.1 Motivação

As viagens aéreas mundiais cresceram em média aproximadamente 5% ao ano nos últimos trinta anos (BELOBABA *et al.*, 2009). Com a demanda de passageiros aumentando, as viagens continuam crescendo regularmente: a taxa de crescimento anual de tráfego aéreo de passageiros em 2017 foi de 8%, e em 2018 foi de 7,4%. Apesar de ter crescido um

pouco menos em 2018 que no ano anterior, continua acima em 2% da taxa de crescimento médio da indústria a longo prazo. De acordo com *International Air Transport Association* - INTERNATIONAL AIR TRANSPORT ASSOCIATION (2019), a previsão é que nas próximas duas décadas a demanda de passageiros seja duplicada. Porém, incorporado ao aumento da demanda, vêm os desafios deste crescimento e dentre eles os aumentos dos atrasos e cancelamentos de voos. De acordo com Rebollo e Balakrishnan (2014), o aumento da demanda diminui a capacidade da rede de absorver interrupções, tornando-a suscetível a atrasos em larga escala.

Os atrasos já são recorrentes e cada vez mais comuns na rotina dos passageiros, principalmente nos aeroportos considerados *hubs* (onde se concentram as conexões). A alta concentração de voos em um aeroporto para a realização de conexões pode gerar atrasos por congestionamento, pois o número de aeronaves tende a se aproximar da capacidade máxima do aeroporto. Os atrasos que podem incorrer, em consequência deste congestionamento, aumentam os custos operacionais às companhias aéreas, o tempo de viagem dos passageiros e criam um trabalho adicional para os controladores de voo, o que aumenta o nível de estresse (WENSVEEN, 2016).

Wensveen (2016) afirma que os atrasos e suas consequências são fatores que causam um impacto negativo na economia. Segundo Baik *et al.* (2010), avaliar o valor econômico de voos atrasados é de interesse tanto de órgãos reguladores quanto de grupos de pesquisa. De acordo com Pyrgiotis *et al.* (2013), uma rede de aeroportos e aeronaves que está intrinsecamente conectada e sobrecarregada possui um grande custo de congestionamento. Este custo compreende custos diretos (companhias aéreas e passageiros) e custos indiretos (indústria aérea e outros setores da economia). Em seu estudo, Baik *et al.* (2010) desenvolveram um método para estimar os custos dos passageiros de voos domésticos incorridos pelos atrasos de voos nos aeroportos dos Estados Unidos. De acordo com Bendinelli *et al.* (2016) os atrasos, além de estressantes aos passageiros e companhias aéreas, custam caro. Os autores buscaram mensurar o impacto dos atrasos de voos nos custos e na dinâmica do transporte aéreo.

Ferguson *et al.* (2013) desenvolveram um estudo onde apresentam os impactos que a economia sofre com os atrasos de voos. As perdas estimadas na economia norte-americana, causadas pelos atrasos de voos em 2007, variaram de US$32.9 bilhões, segundo NEXTOR *et al.* (2010), a US$41 bilhões, conforme o UNITED STATES. Congress. Joint Economic Committee (2008). No mesmo ano, na Europa, foram avaliados os custos estimados pelo atraso no gerenciamento do fluxo de tráfego aéreo, que atingiram mais de US$1,3 bilhão (EUROCONTROL, 2008).

Logo, de acordo com a literatura, a análise dos atrasos aéreos é importante, pois a compreensão de suas potenciais causas pode possibilitar o desenvolvimento de possíveis soluções para ajudar no desempenho do sistema de transporte aéreo (ABDEL-ATY *et al.*,

2007; REBOLLO; BALAKRISHNAN, 2014; SCARPEL; PELICIONI, 2018).

No que diz respeito ao Brasil, com a liberalização do transporte aéreo houve a concentração de voos em alguns aeroportos *hub* (COSTA *et al.*, 2010). De acordo com Wensveen (2016), à medida que o volume de aeronaves se aproxima da capacidade do aeroporto, os atrasos aumentam rapidamente. O Aeroporto Internacional de Guarulhos até o momento é o maior *hub* dos aeroportos brasileiros sendo, assim, o que mais sofre com a concentração de conexões e atrasos por congestionamento (SCARPEL; PELICIONI, 2018).

Nos anuários fornecidos pelo Departamento de Controle do Espaço Aéreo (DECEA) estão disponíveis os *rankings* dos aeroportos brasileiros considerando os movimentos de tráfego aéreo originados da Torre de Controle e Estação e da Estação Aeronáutica. São considerados movimentos de pousos, decolagens, cruzamentos (sobrevoos) e TGL (toques e arremetidas) ocorridos nos aeroportos dentro do período de análise. Seguem na Tabela 1.1 os *rankings* dos sete aeroportos de maior movimento nos anos de 2016 e 2017, fornecidos pelo BRASIL. Departamento de Controle do Espaço Aéreo (DECEA) (2017). Observa-se que o Aeroporto Internacional de Guarulhos (GRU) é o aeroporto mais movimentado do país, sendo a principal porta de entrada no país via aérea e um dos principais *hubs* da América do Sul.

| *Rank* | *Aeroporto* | *2016* | *2017* |
|---|---|---|---|
| 1 | Guarulhos | 272 141 | 271 237 |
| 2 | Congonhas | 219 746 | 223 989 |
| 3 | Brasília | 172 483 | 158 507 |
| 4 | Galeão | 131 168 | 127 092 |
| 5 | Santos Dumont | 120 265 | 118 149 |
| 6 | Campinas | 119 163 | 112 772 |
| 7 | Confins | 100 231 | 100 593 |

TABELA 1.1 – Movimento anual de voos

Desta forma, antecipar a ocorrência de dias congestionados no aeroporto de Guarulhos é de grande importância para o desenvolvimento de estratégias com o propósito reduzir os atrasos e cancelamentos de voos e apoiar seu planejamento. A antecipação de dias congestionados neste trabalho foi feita por meio da análise de uma sequência de dados ao longo do tempo (série temporal), onde foram consideradas as chegadas, as partidas (com atrasos) e os cancelamentos diários de voos no Aeroporto Internacional de Guarulhos. Entre as vantagens de se trabalhar com séries temporais, estão a viabilização de previsões a curto ou longo prazo, descrição de seu comportamento e identificação de periodicidades relevantes nos dados.

## 1.2 Objetivo

O objetivo deste trabalho é a antecipar a ocorrência de dias congestionados no Aeroporto Internacional de Guarulhos (GRU) para apoio ao planejamento. Para alcançar este objetivo, buscou-se a identificação de grupos homogêneos (regimes) dentro da série temporal representada pela fatia diária de voos atrasados e cancelados do Aeroporto Internacional de Guarulhos. A partir dos regimes identificados foi criado um modelo de classificação, comparado em três algoritmos diferentes, para prever com antecedência a ocorrência de dias com grande concentração de voos atrasados e cancelados, podendo assim ser uma ferramenta de suporte a tomada de decisão no aeroporto GRU.

Para que previsões que antecipam a ocorrência de dias com grande concentração de voos atrasados e cancelados sejam feitas, é imprescindível que as variáveis que tenham relevância no impacto de atrasos e cancelamentos de voos sejam detectadas. Scarpel e Pelicioni (2018) concluíram em seu trabalho que as variáveis HHI por *slot* (concentração de mercado), média de *spacing* (a média do intervalo de tempo entre dois movimentos programados de voos), desvio padrão de *spacing* (variação do intervalo de tempo programado), contador de movimentos (movimentos consecutivos do mesmo tipo) e dias da semana são variáveis de impacto, para explicar causas de atrasos, no Aeroporto Internacional de Guarulhos. Assim, consideradas de impacto, estas variáveis foram consideradas no modelo de classificação deste trabalho.

## 1.3 Organização do trabalho

O Capítulo dois apresenta uma revisão de literatura que se inicia com trabalhos relacionados a atrasos e cancelamentos de voos. Segue-se explanando tópicos de séries temporais e detecção de ponto de mudança. Os modelos de Markov também são explorados abordando cadeias Markovianas e dos modelos escondidos de Markov. São apresentados brevemente os métodos de classificação utilizados e, por fim, é apresentado o KDD e justificada sua importância quando há um grande volume de dados.

No Capítulo 3, são apresentados os estágios desenvolvidos neste trabalho. Em um primeiro momento são definidos o escopo deste estudo, é feita a exploração da base de dados bem como o processo de transformação dos dados. Também é definido o tipo de modelo de mineração de dados aplicado. Em um segundo momento são retomadas algumas etapas do processo KDD, onde são definidas variáveis explicativas, são determinados os métodos de classificação utilizados e o critério de avaliação dos modelos. No Capítulo 4, os resultados da aplicação da metodologia são apresentados e discutidos.No Capítulo 5 são apresentadas as conclusões, deste trabalho, e propostas para trabalhos futuros.

# 2 REFERENCIAL TEÓRICO

Neste capítulo são apresentados trabalhos relacionados com o tema, e os métodos empregados no desenvolvimento deste trabalho. Desenvolve-se uma breve explanação de séries temporais e detecção de pontos de mudança, segue-se com os modelos de Markov, estimação de seus parâmetros e critérios de seleção de modelos. São tratados ainda os métodos de classificação bem como a metodologia empregada no trabalho.

## 2.1 Trabalhos Relacionados

Grande parte dos aeroportos de todo o mundo não conseguem operar conforme o planejado devido aos problemas de atraso. Em dias congestionados é alta a proporção de voos afetados por problemas de atrasos. Logo, desenvolver estratégias para reduzir os atrasos é uma questão crítica (SCARPEL; PELICIONI, 2018).

De acordo com Sternberg *et al.* (2017), atrasos são indicadores de desempenho apontados como críticos nos sistemas de transporte aéreo. Atrasos e cancelamentos de voos têm sido analisados por diferentes óticas, como previsão de atraso no taxiamento de voos, conjunto de pequenos fatores influentes nos atrasos, mudança na tendência de demanda, dias congestionados, cancelamentos de voos, capacidade do sistema reduzida e propagação de atrasos (BALAKRISHNA *et al.*, 2010; YU *et al.*, 2019; SCARPEL, 2014; SCARPEL; PELICIONI, 2018; XIONG; HANSEN, 2013; ABDEL-ATY *et al.*, 2007)

No campo da análise de dados, modelos complexos e algoritmos podem ser elaborados utilizando um método de aprendizado de máquina para a realização de análises preditivas. Estes algoritmos procuram detectar padrões e fazer previsões sobre os dados. O uso de algoritmos de aprendizado de máquina tem sido crescente nos últimos anos, devido à complexidade das análises, dado o crescente volume de dados (STERNBERG *et al.*, 2017).

Abdel-Aty *et al.* (2007) aplicaram um método de análise de frequência para detectar padrões periódicos de atrasos de chegadas de voos domésticos no Aeroporto Internacional de Orlando. Os padrões detectados possibilitaram a identificação de variáveis de impacto nas quais foram empregadas técnicas estatísticas para investigar a relação destes fatores

associados aos atrasos. Liu *et al.* (2008) embasaram sua abordagem na geração de cenários utilizando dados históricos da capacidade aeroportuária e desenvolveram um mecanismo de reconhecimento de padrão da capacidade de chegada de voos que pudesse ser amplamente aplicado. O método estatístico de geração de agrupamentos foi empregado para classificar os dados de capacidade de chegada em perfis padrões. Balakrishna *et al.* (2010) modelaram seu problema utilizando um processo Markoviano de decisão e aplicaram um algoritmo de aprendizagem por reforço para prever atrasos no taxiamento de voos que estão partindo.

Buxi e Hansen (2013) produziram cenários de capacidade para auxiliar na tomada de decisões estratégicas diárias no gerenciamento do tráfego aéreo. Os autores desenvolveram três metodologias baseadas em geração de agrupamentos para gerar cenários de capacidade probabilística. Scarpel (2014) desenvolveu um modelo de alerta utilizando árvores de classificação e regressão para criar um modelo de previsão de alerta precoce de mudança na tendência da demanda de aeroportos de São Paulo. O modelo de alerta foi empregado para auxiliar na tomada de decisões gerenciais, a curto e longo prazos, como uma avaliação alternativa com a finalidade de evitar a ocorrência de atraso por congestionamento e apoiar o planejamento da infraestrutura.

Rebollo e Balakrishnan (2014) empregaram abordagens de geração de agrupamento e classificação para a prevenção de atrasos nos horários de partida dos voos em uma ligação específica ou em um aeroporto específico em algum momento no futuro. Os autores compararam a abordagem desenvolvida por eles com modelos de regressão, em aeroportos dos Estados Unidos, fazendo o uso da combinação de variáveis categóricas e explicativas e considerando os âmbitos de previsão de 2, 4, 6 e 24 horas. Chandramouleeswaran *et al.* (2018) apresentaram uma abordagem para prever atrasos nos aeroportos dos Estados Unidos fazendo uma comparação entre redes neurais e regressão. Foram considerados dados temporais, como dia e hora, congestionamentos, redes de aeroportos e o clima. Como parte da abordagem, os autores desenvolveram uma métrica que reduz a dimensionalidade das informações, o que restringe o espaço de atributos e permite o uso dos modelos considerados pelos autores.

Scarpel e Pelicioni (2018) empregaram uma abordagem de análise de dados para construir um modelo de alerta com a finalidade de antecipar a ocorrência de dias congestionados no Aeroporto Internacional de São Paulo. A combinação de abordagens de modelagem, que se baseiam em diferentes premissas, permitiu gerar um modelo com maior flexibilidade e trouxe boas expectativas de melhoria na precisão das previsões. Para a criação do modelo, os autores consideraram a combinação de árvores de classificação e regressão, regressão múltipla e modelos autorregressivos. Yu *et al.* (2019) utilizaram um método de aprendizado não supervisionado combinado com um algoritmo de aprendizado supervisionado de regressão e classificação para realizar análises de prevenção de atrasos

de voos.

Neste trabalho, a análise da série temporal de voos atrasados e cancelados foi feita por meio de detecção de pontos de mudança, sob o conceito de agrupamento. Para que este conceito seja empregado, o método dos Modelos Escondidos de Markov foi utilizado, dada sua ampla aplicação (GHASSEMPOUR *et al.*, 2014; WEI *et al.*, 2015). Os atrativos apresentados por este método, como a simplicidade e o fácil tratamento matemático em estimar os parâmetros do modelo, o torna flexível nas aplicações em séries temporais (ZUCCHINI *et al.*, 2016).

## 2.2 Séries Temporais

As séries temporais são de grande interesse devido sua difusão nas mais diversas áreas como climatologia (GALLAGHER *et al.*, 2013), finanças (TZOURAS *et al.*, 2015), saúde (SALMON *et al.*, 2016), fontes de energias renováveis (LOURENCO *et al.*, 2017), indústria (ASKARI *et al.*, 2016), aviação (BUXI; HANSEN, 2013), entre outros. O histórico do comportamento de uma variável ao longo do tempo projeta informações extremamente úteis dos dados em questão. Por meio de sua análise, é possível investigar o mecanismo gerador da série, fazer previsões a longo ou curto prazo, descrever o comportamento e buscar periodicidades relevantes nos dados (MORETTIN; TOLOI, 2006).

O conjunto de dados de uma série temporal é uma sequência infinita de elementos,

$$S = \{x_1, ..., x_i, ...\} \tag{2.1}$$

onde $x_i$ é um vetor de dados n-dimensionais no instante de tempo $i$ (SHUMWAY; STOFFER, 2017).

Warren (2005) define série temporal como valores que se alteram ao longo do tempo. Aminikhanghahi e Cook (2017) definem séries temporais como uma sequência de medidas ao longo do tempo que descrevem o comportamento de um sistema, ou seja, é uma sequência ordenada de dados obtidos em um intervalo de tempo. Morettin e Toloi (2006) designam séries temporais como parte de uma trajetória de um processo em observação ao longo do tempo. Se a temperatura de um dado local é medida durante o dia, esses dados geram uma série temporal em que o tempo pode ser medido de hora em hora ou de acordo com a necessidade do processo analisado. Ao repetir esta medição por alguns dias, serão geradas séries temporais diárias diferentes, ou de acordo com Morettin e Toloi (2006), trajetórias do processo observado. Assim, os autores definem o conjunto de todas as trajetórias possíveis como um processo estocástico.

A palavra estocástico vem de aleatoriedade. Ao fazer, por exemplo, a análise da série

temporal da Ibovespa, não é possível dizer como ela vai fechar no final do próximo dia. Da mesma forma, não é factível dizer qual será a taxa de desemprego no Brasil no próximo mês. Logo, essa sequência de variáveis aleatórias ordenadas no tempo, denominada série temporal, é um processo estocástico. Wooldridge (2016) reitera que, ao analisar uma série temporal, é possível obter uma saída, ou uma realização do processo estocástico. Só é viável alcançar uma única realização porque não se pode voltar no tempo e começar o processo novamente.

As séries temporais podem ser discretas (obtidas em intervalos de tempo discretos e equidistantes) ou contínuas (em qualquer intervalo de tempo). Sua periodicidade pode variar de acordo com a necessidade do fenômeno analisado. As séries temporais podem ainda ser classificadas como estacionárias e não-estacionárias. Wooldridge (2016) define que o processo de uma série temporal estacionária é aquele no qual a distribuição de probabilidade é estável ao longo do tempo. Se há um conjunto de variáveis aleatórias em sequência e em seguida altera-se essa sequência de $h$ períodos à frente, a distribuição de probabilidade deve permanecer inalterada.

O comportamento de uma série temporal pode sofrer mudanças ao longo do tempo, uma vez que os parâmetros da distribuição referentes a ela, repentinamente, podem se alterar. O processo de identificar essas mudanças ao longo do tempo é denominado Detecção de Pontos de Mudança (*Change point detection – CPD*). Conceitos como de segmentação, detecção de borda, detecção de eventos, estimativas de pontos de mudança e detecção de anomalia apresentam estreita relação com o conceito de pontos de mudança (AMINIKHANGHAHI; COOK, 2017). Estes conceitos não necessariamente trabalham em torno de séries temporais. O que apresentam em comum é a identificação de mudanças ou anomalias ocorridas no processo analisado.

O problema de detecção de pontos de mudança pode apresentar dois objetivos diferentes: identificar se há mudança (teste de hipóteses), e localizar um ou mais pontos de mudança presentes (problema de estimação) (CHEN; GUPTA, 2012). Dentre as mais variadas vertentes de se tratar o problema de CPD, Aminikhanghahi e Cook (2017) afirmam que, por uma perspectiva diferente, este pode ser considerado um problema de geração de agrupamentos. Neste caso, o objetivo é identificar a estrutura em um conjunto de dados, organizando-os em grupos homogêneos, nos quais a similaridade dentro do grupo é minimizada e a dissimilaridade entre os grupos é maximizada (WARREN, 2005).

## 2.3 Detecção de Pontos de Mudança

Detecção de Pontos de Mudança é a estimação de pontos em uma série temporal para os quais as propriedades estatísticas são diferenciadas e os intervalos entre os pontos en-

contrados são chamados de regimes ou estados, representando um conjunto de observações com características comuns entre si. De acordo com Killic e Eckley (2014), detecção de pontos de mudança é a determinação de pontos nos quais as propriedades estatísticas são alteradas em uma sequência de observações. Aminikhanghahi e Cook (2017) apontam CPD como variações súbitas em uma série temporal que podem representar uma transição entre regimes. James e Matteson (2014) afirmam que é um processo no qual se detecta mudança de distribuição em uma sequência de observações ordenadas no tempo.

Considerando observações nos instantes de tempo $t$ e $t+1$, se um ponto no instante $t$ pertence a um grupo diferente de uma observação no instante $t+1$, então um ponto de mudança ocorre entre as duas observações. Pela perspectiva de CPD como um problema de geração de agrupamento, as observações que pertencem à série temporal, localizadas entre um ponto de mudança e outro, formam conjuntos de observações que compartilham das mesmas propriedades estatísticas, denominados neste trabalho como regimes, como mostra a Figura 2.1. Nela, pode-se observar dois pontos de mudança (onde as propriedades estatísticas se alteram) e três regimes diferentes detectados (segmentos em que as propriedades estatísticas são comuns aos conjuntos de observações).

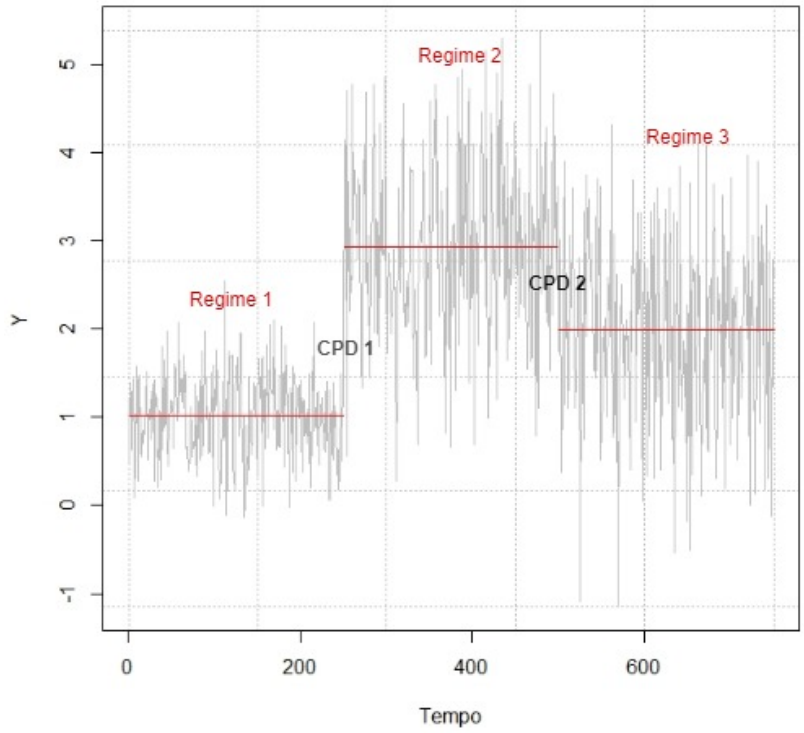



FIGURA 2.1 – Pontos de mudança e regimes detectados em uma série temporal

O conceito de CPD começa a emergir entre as décadas de 1920 e 1930 motivado pelas pesquisas de Shewhart (1926) na área de controle de qualidade, o que contribuiu para o

surgimento de uma vasta gama de pesquisas (TARTAKOVSKY *et al.*, 2015). Ainda em conexão com o controle de qualidade industrial, na década de 1950 surgem as primeiras publicações de Page (1954) e Girshick e Rubin (1952), mas somente no final dos anos 1950 há uma formulação matemática feita por A.N.Kolmogorov (BRODSKY; DARKHOVSKY, 1993).

Aminikhanghahi e Cook (2017) ressaltam que CPD foi intensamente estudado nos campos de mineração de dados, estatística e ciência da computação nas últimas décadas, abrangendo diversas áreas de aplicação. O monitoramento de condições médicas como batimento cardíaco e eletroencefalograma, reconhecimento de voz, análise de atividades humanas, climatologia e finanças, são alguns exemplos dos inúmeros setores em que CPD é empregado.

Existem inúmeras maneiras de tratar o problema de ponto de mudança, dentre eles a mudança na média, mudança na variância, mudança na média e na variância, mudança na média do vetor e mudança na covariância. Os principais métodos para detectar pontos de mudança são: (i) razão de verossimilhança, (ii) não paramétricos e (iii) Abordagem Bayesiana (CHEN; GUPTA, 2012).

Os algoritmos utilizados para identificar pontos de mudança são classificados como online e offline. Os algoritmos online são utilizados em tempo real, no qual se considera certo número de observações em um processo de monitoramento com o intuito de identificar novos pontos de mudança (AMINIKHANGHAHI; COOK, 2017). Atualmente estão ganhando espaço em diferentes áreas, como na avaliação de desempenho de hardware (MISHIN *et al.*, 2014), análise de redes sociais (APREM; KRISHNAMURTHY, 2017), e estão fortemente presentes na área de controle estatístico de qualidade, vigilância sanitária e processamento de sinais (MEI, 2006). Pesquisas recentes têm sido realizadas, com o intuito de diminuir os pontos de mudanças que atestam falsos positivos. Nos algoritmos *offline* são considerados os históricos de todo conjunto de dados para identificar onde ocorreram mudanças.

### 2.3.1 Formulação do problema

O ponto de mudança configura a transição entre os diferentes regimes que compõe a série temporal. A transição representa as propriedades estatísticas que são distintas entre os regimes.

Para a identificação de um ponto de mudança considere $Y_{m:n} = Y_m$, $Y_{m+1}$, ..., $Y_n$ uma sequência de variáveis de uma série temporal, o instante de tempo $t \in \{m, ...,\text{n}-1\}$; um ponto de mudança ocorre quando as propriedades estatísticas de $\{Y_m, ..., Y_t\}$ e $\{Y_{t+1}, ..., Y_n\}$ são diferentes.

Para determinar múltiplos pontos de mudança considerando a série temporal $Y_{m:n}$, duas hipóteses são testadas:

Hipótese nula ($H_0$), onde não ocorre mudança,

$$H_0 : P_{Y_m} = ... = P_{Y_k} = ... = P_{Y_n} \tag{2.2}$$

versus hipótese alternativa ($H_A$), na qual ocorrem mudanças, isto é, $\exists m < k_* < n$, $k_* = \{k_1, k_2, ..., k_q\}$, tal que

$$H_A : P_{Y_m} = ... = P_{Y_{k_*}} \neq P_{Y_{k_*+1}} = ...P_{Y_{k_*q}} \neq P_{Y_{k_*q+1}}... = P_{Y_n}, \tag{2.3}$$

onde $P$ é uma distribuição de probabilidade, $q$ o número de pontos de mudança e $k_*$ as posições desconhecidas dos pontos de mudança.

### 2.3.2 Classificações dos problemas de pontos de mudança

Algumas das classificações utilizadas para os problemas de CPD são propostas por Brodsky e Darkhovsky (1993), a partir de amostras não homogêneas de dados, mas que podem conter segmentos de homogeneidade.

#### 2.3.2.1 Aquisição de dados

No método da aquisição de dados, existem duas possibilidades: (i) quando é feita a aquisição de toda a base de dados para checar a hipótese de homogeneidade ($H_0$), problema denominado *a posteriori change point problem*; (ii) quando a hipótese é checada simultaneamente com as observações, ou seja, *online* juntamente com a aquisição dos dados (esse problema é conhecido como *sequential change point problem*).

#### 2.3.2.2 Completude da informação estatística a *priori*

De acordo com informação da estatística a *priori*, os problemas de CPD podem ser distinguidos entre (i) métodos paramétricos; (ii) semi-paramétricos; e (iii) não-paramétricos. Os métodos paramétricos são baseados nas informações a *priori* completas, ou seja, um modelo probabilístico.

Os métodos não-paramétricos possuem uma quantidade mínima de informações estatísticas a *priori* necessárias. Estas informações consistem na suposição de que algumas características probabilísticas das observações (esperança, dispersão, função de correlação, densidade do espectro), estão sendo alteradas em momentos desordenados. Entre métodos

paramétricos e não-paramétricos há uma vasta área para os métodos semi-paramétricos. De modo geral, supõe-se que a função de distribuição de probabilidade das observações pertence a alguma classe de função de distribuição e os parâmetros desta distribuição se alteram em momentos desordenados.

#### 2.3.2.3 Características dos dados

De acordo com as características dos dados, duas distinções podem ser feitas: (i) problemas de mudança de processos aleatórios; e (ii) problemas de mudança de campos aleatórios. Uma mudança (*change point*) de um processo aleatório é considerada por Brodsky e Darkhovsky (1993) como um momento do tempo em que algumas características probabilísticas deste processo se alteram. Uma mudança de campos aleatórios é uma zona particular no domínio do campo que difere quanto a algumas características probabilísticas das observações.

Ainda, segundo os autores, quanto à dependência estatística entre as observações, podem ser formulados problemas de mudança para sequências aleatórias com observações independentes, e problemas de mudança com variáveis aleatórias dependentes. A dependência estatística entre as observações pode ser descrita em termos de condições de misturas para um processo aleatório pelos modelos de Markov.

Quanto aos tipos de mudanças, alguns autores consideram, além de mudanças abruptas, mudanças graduais nas características probabilísticas. Dependendo do número de pontos de mudança, um problema de mudança pode ser distinguido entre uma ou múltiplas desordens. Um ponto de mudança é geralmente considerado um momento determinístico desconhecido, ou uma variável aleatória com uma distribuição a *priori* conhecida. Além das classificações citadas, existem outras propostas de classificações dos problemas de pontos de mudança (BRODSKY; DARKHOVSKY, 1993).

## 2.4 Modelos de Markov

Para tratar o problema de CPD, visto como um problema formação de agrupamentos, foi empregado o Modelo Escondido de Markov (*Hidden Markov Models – HMM* ), em que o objetivo é a identificação de estados (GHASSEMPOUR *et al.*, 2014; WEI *et al.*, 2015). Luong *et al.* (2012) afirmam que os dados são as observações do HMM, e os segmentos desconhecidos os estados ocultos.

São apresentadas nas próximas seções a formulação do Modelo Escondido de Markov, suas propriedades e premissas.

### 2.4.1 Cadeias de Markov

De acordo com Zucchini *et al.* (2016), as cadeias de Markov (*Markov Chains* - MC) supõem a dependência das observações no instante de tempo futuro $t+1$, nas observações atuais no instante de tempo $t$. Estas observações são uma sequência de variáveis aleatórias definidas como estados. Logo, considere que a sequência de variáveis aleatórias $\{S_t : t \in \mathbb{N}\}$ representa uma Cadeia de Markov se, para todo instante de tempo $t$, a seguinte propriedade é satisfeita:

$$P(S_{t+1}|S_{1:t}) = P(S_{t+1}|S_t), \tag{2.4}$$

na qual dizer que a sequência histórica $S_{1:t} = \{S_1, S_2, ..., S_t\}$ condiciona $S_{t+1}$, é equivalente a dizer que somente o estado presente $S_t$ condiciona o estado futuro $S_{t+1}$, logo $S_{t+1}$ depende somente de $S_t$. A Figura 2.2 ilustra como cada estado, tanto no passado como no futuro, depende somente do seu estado imediatamente anterior.


FIGURA 2.2 – Cadeia de Markov - adaptada de Zucchini *et al.* (2016)

Importantes quantidades associadas a MC são as probabilidades condicionais, denominadas Probabilidades de Transição, onde

$$P(S_{c+t} = j|S_c = i) \tag{2.5}$$

onde o estado $S_{c+t}$ representa a probabilidade de $S_c$ no instante $t$ de ir ao estado $j$ dado $S_c$ estar no estado anterior $i$. As probabilidades de transição são denotadas por:

$$\lambda_{ij} = P(S_{c+t} = j|S_c = i). \tag{2.6}$$

A matriz de probabilidades de transição, conhecida também por matriz de transição, pode ser definida como uma matriz quadrada $(i, j)$ contendo elementos $\lambda_{ij}$.

$$\Lambda = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1m} \\ \vdots & \ddots & \vdots \\ \lambda_{m1} & \cdots & \lambda_{mm} \end{pmatrix}$$

em que, $\sum_{j=1}^{n} \lambda ij = 1$, ou seja, as linhas de $\Lambda$ somam 1.

Além das probabilidades condicionais, há a probabilidade incondicional $P(S_t = j)$, definida como a probabilidade de se manter em um determinado estado em um instante de tempo $t$. A probabilidade incondicional é denotada por um vetor de linhas na forma

$$u(t) = (P(S_t = 1), ..., P(S_t = m)), t \in \mathbb{N}, \tag{2.7}$$

em que $u(1)$ é a distribuição inicial da MC.

### 2.4.2 Modelos Escondidos de Markov

Os Modelos Escondidos de Markov (Hidden Markov Models – HMM) são modelos nos quais a distribuição de probabilidade que gera uma observação depende de um estado pertencente ao processo oculto de Markov (ZUCCHINI *et al.*, 2016). De acordo com Visser (2011) os estados escondidos de um HMM correspondem a distribuições de probabilidade cujo número de estados e os parâmetros são desconhecidos.

Os atrativos apresentados pelo HMM, como simplicidade e o fácil tratamento matemático para calcular os parâmetros do modelo, o torna flexível para aplicações em séries temporais. *Latent Morkov model* é frequentemente aplicado em diferentes áreas de ciências sociais. Em psicologia, por exemplo, são utilizados para modelar processos de aprendizado. Em economia, são chamados de modelos de mudança de regimes, outras aplicações, como reconhecimento de fala, *latent Morkov model* normalmente é denominado *hidden Morkov model* (VISSER; SPEEKENBRINK, 2010).

De acordo com Zucchini *et al.* (2016), nos últimos anos o HMM foi muito aplicado no processo de reconhecimento automático de voz e tem se expandindo para outros campos como todos os tipos de reconhecimento (face, gestos, manuscritos, assinatura), bioinformática (análises de sequências biológicas), meio ambiente (direção do vento, precipitação, terremotos), finanças (séries de retornos diários), biofísica (modelagem de canais iônicos) e ecologia (comportamento animal).

Zucchini *et al.* (2016) descrevem o modelo escondido de Markov $O_t : t \in \mathbb{N}$ como um tipo de mistura dependente, em que as sequências históricas das observações $O_{1:t}$ e dos estados escondidos $S_{1:t}$ são representadas por dois processos definidos como,

$$P(S_t|S_{1:(t-1)}) = P(S_t|S_{t-1}), \quad t = 2, 3, ... \tag{2.8}$$

$$P(O_t|O_{1:(t-1)}, S_{1:t}) = P(O_t|S_t), \quad t \in \mathbb{N}, \tag{2.9}$$

em que a Equação 2.8 representa um processo de parâmetros (*parameter process*) não

observados $S_t : t = 1, 2, ...$, que satisfaz a propriedade de Markov. A Equação 2.9 representa um processo dependente de estados de forma que a distribuição da observação no instante t $O_t : t = 1, 2, ...$ depende somente do estado atual e não de estados ou observações anteriores. Se a MC $(S_t)$ tem $n$ estados, $O_t$ será um HMM de $n$ estados. A Figura 2.3 configura a estrutura dos modelos escondidos de *Markov*, em que os estados escondidos, representados por $S_n$, dependem somente do estado imediatamente anterior, e os estados observáveis, representados por $O_n$, dependem somente do estado escondido corrente.

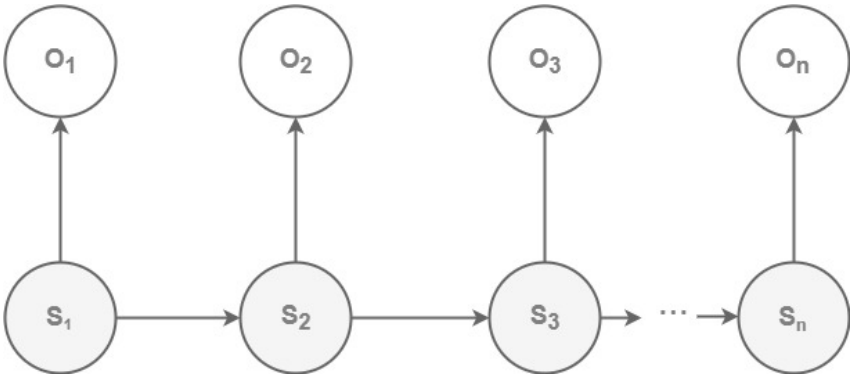



FIGURA 2.3 – Modelos Escondidos de Markov - adaptada Zucchini *et al.* (2016)

A notação a seguir descreve tanto observações discretas quanto contínuas. Caso sejam discretas, são definidas por $i = 1, 2, ..., n$,

$$p_i(o) = P(O_t = o|S_t = i), \tag{2.10}$$

em que $p_i$ é a função de probabilidade de $O_t$, se MC estiver em um estado $i$ no instante de tempo $t$. No caso de uma função contínua, $p_i$ é definida como função densidade de probabilidade associada com o estado $i$. *State-dependent distributions* faz referência as $n$ distribuições $p_i$ do modelo.

A distribuição marginal de um conjunto de observações é uma distribuição de mistura cujos dados são extraídos de duas ou mais distribuições com diferentes valores de parâmetros (VISSER, 2011). De acordo com o autor, outra forma de afirmar isto é dizer que os HMMs têm estados discretos que geram os dados. Considerar HMM com distribuição discreta implica que $S_T$ são elementos de um conjunto finito $i = 1, ..., n$ de forma que $S_t = k, k \in i$, em que n é o número de estados do modelo. Logo, a sequência de variáveis dos estados $S_{1:T}$ são variáveis aleatórias com valores discretos. Dado que $i$ é um conjunto finito, a distribuição marginal dos dados é uma distribuição de mistura com $k$ estados:

$$f(O_T) = \sum_{k=1}^{n} p_k f_k(O_t), \tag{2.11}$$

em que $p_k$ são as proporções de estados, com a restrição $\sum_{k=1}^{n} p_k = 1, p_k \geq 0$ e $f_k(.)$ é a distribuição condicional no estado $k$ .

Sintetizando, os modelos escondidos de Markov são definidos como modelos com estados discretos, caracterizados por suas funções de distribuição onde a evolução dos estados no decorrer do tempo é governada por um processo de Markov (VISSER, 2011).

### 2.4.3 Verossimilhança

Verossimilhança é uma medida proporcional a uma probabilidade dada por $L(H) = KP(D|H)$, em que a verossimilhança da hipótese $(H)$, dadas as observações $(D)$, é proporcional à probabilidade de obter $D$ dado que $H$ seja verdadeiro, multiplicado por uma constante arbitrária $K$ (ETZ, 2018).

Zucchini *et al.* (2016) denotam $L_T$ como a verossimilhança em que $T$ são consideradas observações consecutivas $y_1, y_2, ..., y_T$ geradas por um HMM. O que se pretende é buscar a probabilidade $L_T$ de observar esta sequência, calculada sob um estado $n$ de HMM que possui distribuição inicial $\phi$, matriz de transição $\Lambda$ para MC e funções de probabilidade (densidade) dependentes do estado $p_i$. A verossimilhança é dada por,

$$L_T = \phi P(y_1)\Lambda P(y_2)\Lambda P(y_3)...\Lambda P(y_T)1' \tag{2.12}$$

Se $\phi$, a distribuição do estado inicial da cadeia de Markov $S_1$, é estacionária, tem-se:

$$L_T = \phi\Lambda P(y_1)\Lambda P(y_2)\Lambda P(y_3)...\Lambda P(y_T)1' \tag{2.13}$$

Para formular o algoritmo *forward*, define-se o vetor $\alpha_t$, para $t = 1, 2, ..., T$, por

$$\alpha_t = \phi P(y_1)\Lambda P(y_2)\Lambda P(y_3)...\Lambda P(y_t) = \phi P(y_1)\prod_{s=2}^{t}\Lambda P(y_s), \tag{2.14}$$

onde o produto vazio é a matriz identidade. Segue-se desta definição que:

$$L_T = \alpha_T 1' \tag{2.15}$$

e $\alpha_t = \alpha_{t-1}\Lambda P(y_t)$ para $t \geq 2$.

Assim, de forma mais conveniente, a Equação 2.15 mostra os cálculos envolvidos na Equação 2.12 para encontrar $L_T$:

$\alpha_1 = \phi P(y_1)$

$\alpha_t = \alpha_{t-1}\Lambda P(y_t), \qquad t = 2, 3, ..., T$

E em caso estacionário,

$\alpha_0 = \phi$

$\alpha_t = \alpha_{t-1}\Lambda P(y_t), \qquad t = 1, 2, ..., T,$

que correspondem, de modo mais conveniente, aos componetes da Equação 2.13 utilizados para encontrar o valor de $L_T$ da Equação 2.15.

Para uma distribuição de mistura, os parâmetros são geralmente estimando por meio do método da máxima verossimilhança (*maximum likelihood estimation - MLE*), em que o objetivo é encontrar a probabilidade de maior valor em um conjunto de parâmetros, considerando o contexto em que os dados são não observados ou faltantes. Para se obter a estimativa de máxima verossimilhança dos parâmetros do modelo, é necessário determinar a verossimilhança marginal das observações; para tanto, um dos algoritmos mais utilizados é o *forward-backward algorithm* ou sua variante *forward* (VISSER; SPEEKENBRINK, 2010). Uma vantagem importante da expressão matricial para a verossimilhança é o *Forward algorithm* para o cálculo recursivo da verossimilhança. Para maiores detalhes consulte (ZUCCHINI *et al.*, 2016).

Lystig e Hughes (2002) adaptaram o algoritmo *forward*, permitindo o cálculo dos gradientes do logaritmo da verossimilhança (*log-likelihood* - LL) ao mesmo tempo. Para isso, a verossimilhança foi reescrita como segue:

$$L_T = P(O_{1:T}) = \prod_{t=1}^{T} P(O_t|O_{t-1}), \tag{2.16}$$

onde $P(O_1|O_0) = P(O_1)$. O logaritmo de verossimilhança pode ser escrito como,

$$l_T = \sum_{t=1}^{T} log[P(O_t|O_{1:(t-1))}]. \tag{2.17}$$

Para o cálculo do logaritmo da verossimilhança, é utilizado um algoritmo recursivo (*forward*) para as probabilidades condicionais, começando pela probabilidade conjunta do processo observado e do estado não observado do processo escondido no primeiro ponto.

$$\gamma_1(j) = P(O_1, S_1 = j) = \pi_j b_j(O_1) \tag{2.18}$$

$$\gamma_t(j) = P(O_t, S_t = j|O_{1:(t-1)}) = \sum_{i=1}^{N} [\gamma_{t-1}(i) a_{ij} b_j(O_t)](\Gamma_{t-1})^{-1}, \tag{2.19}$$

em que $\Gamma_t = \sum_{i=1}^{N} \gamma_t(i)$. Combinando $\Gamma_t = P(O_t|O_{1:(t-1)})$, e a Equação 2.17, o

logaritmo da verossimilhança pode ser escrito como:

$$l_T = \sum_{t=1}^{T} log\Gamma_t. \tag{2.20}$$

### 2.4.4 Estimação de parâmetros

Um método utilizado para estimar parâmetros é o algoritmo de maximização da esperança (*expectation-maximization* - EM). Como a sequência dos estados ocupados pela MC de um HMM não é observável, uma forma de estimar os parâmetros é tratar estes estados como faltantes e empregar o algoritmo EM para encontrar a estimativa de MLE dos parâmetros (ZUCCHINI *et al.*, 2016). No algoritmo EM, os parâmetros são estimados maximizando iterativamente o LL conjunto esperado dos parâmetros, dadas as observações e os estados (VISSER; SPEEKENBRINK, 2010).

De maneira geral, o algoritmo pode ser descrito em dois passos descritos a seguir: (i) o passo E, que calcula as esperanças condicionais dos estados dadas as observações e dada a estimativa corrente dos parâmetros (em outras palavras as esperanças condicionais são calculadas a partir dos dados observados e um valor inicial dos parâmetros não observados); (ii) o passo M que maximiza o LL dos dados faltantes e observados, utilizando as esperanças condicionais encontradas no passo E. Assim, os valores encontrados são utilizados em outro passo E, e o algoritmo repete os passos até alcançar um critério de convergência (ZUCCHINI *et al.*, 2016).

Seja $\theta = (\theta_1, \theta_2, \theta_3)$ o vetor dos parâmetros composto por três sub-vetores. De acordo com Visser e Speekenbrink (2010), o LL conjunto pode ser escrito como,

$$\begin{aligned} logP(O_{1:T}, S_{1:T}|z_{1:T}, \theta) = logP(S_1|z_1, \theta_1) + \sum_{t=2}^{T} logP(S_t|S_{t-1}, z_{t-1}, \theta_2) \\ + \sum_{t=1}^{T} logP(O_t|S_t, z_t, \theta_3) \end{aligned} \tag{2.21}$$

Onde a verossimilhança depende dos estados não observados $S_{1:T}$. No passo E, os estados não observados são substituídos por seus valores esperados, dado o conjunto de parâmetros iniciais $\theta' = (\theta'_1, \theta'_2, \theta'_3)$ e as observações $O_{1:T}$. O logaritmo da verossimilhança esperado,

$$Q(\theta, \theta') = E_{\theta'}(logP(O_{1:T}, S_{1:T}|O_{1:T}, z_{1:T}, \theta)), \tag{2.22}$$

pode ser escrito como,

$$Q(\theta,\theta') = \sum_{j=1}^{n} \iota_1(j) log P(S_1 = j|z_1,\theta_1) + \sum_{t=2}^{T}\sum_{j=1}^{n}\sum_{k=1}^{n} \varepsilon_t(j,k) log P(S_t = k|S_{t-1} = j, z_{t-1},\theta_2) + \sum_{t=1}^{T}\sum_{j=1}^{n}\sum_{k=1}^{m} \iota_t(j) log P(O_t^k|S_t = j, z_t,\theta_3), \quad (2.23)$$

onde os valores esperados $\varepsilon_t(j,k) = P(S_t = k, S_{t-1} = j|O_{1:T}, z_{1:t},\theta')$ e $\iota_t(j) = P(S_t = j|O_{1:T}, z_{1:T},\theta')$ podem ser calculados por meio do algoritmo *forward-backward*. O passo M consiste em maximizar a Equação 2.23 em função do parâmetro $\theta$, podendo maximizar separadamente $\theta = (\theta_1, \theta_2,$ e $\theta_3)$. Observe que, para uma maximização futura, os valores esperados $\iota_t(j)$ são utilizados como pesos iniciais para observações de $O_t^k$ (VISSER; SPEEKENBRINK, 2010).

### 2.4.5 Critérios de seleção de modelos

Em um HMM, Zucchini *et al.* (2016) afirmam que o modelo é melhor ajustado à medida que se aumenta o número de estados (critério *likelihood*). Entretanto, o aumento do número de estados aumenta exponencialmente o número de parâmetros, o que torna o problema muito complexo. De acordo com os autores, para melhor ajustar o modelo deve-se considerar ao aumento do número de parâmetros.

Dois critérios populares de seleção do melhor ajuste de modelo são *Akaike Information Criterion* (AIC) e *Bayesian Information Criterion* (BIC). Os dois critérios são baseados na função de máxima verossimilhança, e a diferença entre ambos está na forma de penalização dos modelos. Com relação ao número de parâmetros: AIC penaliza menos o modelo com mais parâmetros que BIC (KUHA, 2004). Normalmente os modelos que apresentam menores valores de AIC e BIC são considerados os de melhor ajuste.

Ainda de acordo com Kuha (2004), a abordagem Bayesiana que motiva o BIC identifica os modelos nos quais há a probabilidade de serem reais, e destes assume um como verdadeiro. Entretanto, o AIC nega explicitamente a existência de um modelo verdadeiro identificável e usa a previsão esperada de dados futuros como o critério de adequação de um modelo.

Os critérios AIC e BIC são definidos como,

$$AIC = -2logL + 2p \quad (2.24)$$

$$BIC = -2logL + plogN \tag{2.25}$$

Em que,
$L$ = *Log-likelihood* do modelo ajustado;
$p$ = Número de parâmetros do modelo;
$N$ = Número de observações.

Kuha (2004) afirma que os critérios AIC e BIC normalmente identificam bons modelos para dados observados, mas ambos os critérios ainda podem falhar. O autor sugere o uso dos dois critérios juntos, e afirma que quando os critérios concordam em relação ao melhor modelo, há garantia da robustez da escolha.

## 2.5 Métodos de Classificação

O HMM identifica diferentes estados observados, que podem ser vistos como grupos imbuídos de atributos próprios. Assim, se faz necessário o uso de um classificador para que, por meio da seleção de variáveis, seja criado um modelo de previsão. Com o propósito de uma análise comparativa, são considerados três métodos: (i) Árvores de Classificação e Regressão (*Classification and Regression Tree - CART*); (ii) Florestas Aleatórias (*Random Forests* - RF); e (iii) Máquinas de Vetores de Suporte (*Support Vector Machine* - SVM). Neste trabalho foi feita uma breve apresentação destes métodos de classificação. Para maiores informações consultar (JAMES *et al.*, 2013; HASTIE *et al.*, 2009).

### 2.5.1 Árvores de Classificação e Regressão

Árvores de classificação e regressão são conceitualmente simples, úteis para interpretação e visualização podendo ser utilizadas para ambos, regressão e classificação. As variáveis respostas das árvores de regressão são dadas pela resposta média das observações de treino que pertencem ao mesmo nó. No entanto, as árvores de classificação preveem que cada observação pertence à classe de observações de treinamento mais comum na região à qual pertence (JAMES *et al.*, 2013).

De acordo com Scarpel (2014), CART é atraente quando a interpretação é uma questão importante, uma vez que os dados são projetados para detectar as variáveis de predição importantes e gerar uma estrutura de árvore para representar a partição identificada. O algoritmo de uma árvore de regressão pode ser resumido em quatro etapas:

1. Fazer partições binárias recursivas para expandir a árvore, parando apenas quando cada nó terminal tiver menos que o número mínimo de observações;

2. Aplicar a poda em função da complexidade de custos em árvores grandes, para obter uma sequência das melhores subárvores, em função de um parâmetro $\alpha$;

3. Utilizar a validação cruzada *K-fold* para escolher $\alpha$. Para cada $k = 1, ..., K$; a) Repetir os passos 1 e 2 em todos, exceto na k-ésima partição dos dados de treinamento; b) Avaliar o erro quadrático médio previsto na k-ésima partição deixada de fora em função de $\alpha$; Fazer a média dos resultados para cada valor de $\alpha$ e escolher $\alpha$ para minimizar o erro médio;

4. Retornar a subárvore da etapa 2 que corresponde ao valor escolhido de $\alpha$.

A expansão de uma árvore de classificação é bastante similar. Entretanto, não se pode utilizar soma de quadrados dos resíduos como critério para partições. A taxa de erro de classificação é feita por meio das medidas *Gini* e Entropia. Para que se determine o tamanho ideal de uma árvore, um método considerado é a regra de "um desvio padrão". Por esse método, se escolhe a menor árvore cujo erro relativo de validação cruzada seja próximo ao erro relativo de validação cruzada mínimo mais um desvio padrão (SCARPEL, 2014).

### 2.5.2 Florestas Aleatórias

Florestas Aleatórias é um método de classificação composto por várias árvores de decisão, que combina o conceito de *bagging* com a seleção aleatória de variáveis a cada partição. De acordo Breiman (2001), RF consiste em usar um subconjunto de variáveis de entrada, selecionadas aleatoriamente para expandir cada árvore. Dentre os benefícios de RF inclui-se boa acurácia, relativa robustez a *outliers* e ruídos, além de fornecer estimativas internas úteis de erro, e avaliação da importância relativa das variáveis.

O autor concluiu que RF é uma ferramenta eficaz para fazer previsões. Injetar o tipo certo de aleatoriedade a torna um método de classificação e regressão preciso. A estrutura, em termos de força dos preditores individuais, e suas correlações fornecem *insights* sobre a capacidade da floresta aleatória em realizar previsões.

Kandhasamy e Balamurali (2015) apresentam a expansão das árvores que compõem o método RF como segue:

1. Se o número de observações no conjunto de treinamento é $N$, faz-se a amostragem de $N$ observações aleatoriamente, com reposição, a partir dos dados originais. As amostras serão o conjunto de treinamento;

2. Se houver $M$ variáveis de entrada, um subconjunto delas é selecionado aleatoriamente e a melhor partição é utilizada para dividir o nó. O valor de $M$ é mantido constante durante a expansão da floresta;

3. Cada árvore deve se expandir o máximo possível, e não há podas.

### 2.5.3 Máquinas de Vetores de Suporte

De acordo com Lorena e De Carvalho (2007), SVM é baseado na Teoria de Aprendizado Estatístico, na qual princípios devem ser seguidos na obtenção de uma boa generalização (capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu). Sua formulação incorpora o princípio da minimização de risco estrutural e o princípio da minimização do risco empírico. A aplicação da Teoria de Aprendizado Estatístico assume que os dados são gerados independente e identicamente distribuídos. O risco estrutural mede a capacidade de generalização de um classificador. O risco empírico mede o desempenho do classificador nos dados de treinamento, de acordo com os dados classificados incorretamente.

Segundo Karatzoglou *et al.* (2006), o SVM pode ser visto como um algoritmo linear em um espaço de dimensão $n$. Como $n$ pode ser um número alto de dimensões, o SVM oferece duas vantagens destacáveis: é um método baseado em funções *kernel* (por exemplo, funções de base radial e funções polinomiais), o que possibilita de trabalhar em espaços de qualquer dimensão sem custo computacional significativo; e a possibilidade de utilizar uma função *kernel* para um problema que pode ser aplicado diretamente na base de dados sem a necessidade de redução dimensional por meio da extração de características.

O algoritmo de SVM pode ser generalizado de casos lineares para não lineares. De modo geral, quando a base de dados não pode ser linearmente separada, uma maneira de se tratar é por meio do custo (parâmetro que controla a penalidade paga pelo SVM por classificações errôneas). Um alto valor do custo forçará o SVM a criar uma função de previsão complexa, classificando erroneamente os dados de treino o mínimo possível (KARATZOGLOU *et al.*, 2006).

Para problemas de classificação, o SVM pode ser considerado binário, em que as classes são separadas por hiperplanos (lineares ou não lineares). O objetivo é separar as observações em dois grupos, utilizando uma função que seja capaz de separar futuras observações com uma certa precisão. Para se utilizar o o SVM, em que se tem mais de duas classes, foram propostas algumas abordagens: (i) um-contra-um, em que são gerados SVMs binários e todos os pares de classes são comparados, considera-se a classe à qual foi atribuída com maior frequência; e (ii) um-contra-todos, onde são gerados SVMs em que cada classe é comparada às demais e uma classe é gerada como resultado. Considera-se a classe com maior número de indicações (JAMES *et al.*, 2013).

## 2.6 Descoberta de Conhecimento em Base de Dados

Um número muito grande de registros digitais gera base de dados volumosas. Lidar, tratar as bases de dados para extrair informações úteis é de grande valor, pois a interpretação destes registros pode fornecer conhecimentos valiosos. Fayyad *et al.* (1996b) afirmam que estes fluxos de registros digitais arquivados em grandes bancos de dados normalmente são chamados de armazéns de dados. Para tratar e extrair conhecimento destes armazéns de dados surgiram técnicas e ferramentas, que fazem parte da descoberta de conhecimento em base de dados (*Knowledge Discovery in Databases* - KDD) e mineração de dados (*data mining* - DM).

O termo KDD surgiu em 1989 para enfatizar que o conhecimento é o produto final da descoberta orientada por dados e refere-se ao amplo processo de descoberta de conhecimento da base de dados. DM é considerada uma etapa particular dentro do processo KDD, onde é feita a aplicação de algoritmos específicos para a extração de padrões da base de dados. O termo DM tem sido utilizado por estatísticos, analistas de dados e pela comunidade de gerenciamento de sistemas de informação, enquanto o termo KDD tem sido mais usado na área de inteligência artificial e aprendizado de máquina (FAYYAD *et al.*, 1996a).

### 2.6.1 O processo KDD

O processo KDD começa com a determinação dos objetivos, e termina com a implementação da descoberta do conhecimento. É o processo em que um conjunto de dados passa pela seleção, pré-processamento e transformações necessárias, em seguida são aplicados métodos de mineração de dados em que se almeja encontrar padrões, avaliar os resultados e identificar subconjuntos destes padrões (conhecimento resultante de todo o processo).

Maimon e Rokach (2010) elucidam o processo de descoberta de conhecimento como iterativo e interativo, consistindo de nove etapas, onde o processo é iterativo em cada etapa, o que significa que a movimentação para trás, para ajustar às etapas anteriores, pode ser necessária. O processo tem muitos aspectos peculiares. Não é possível apresentar uma fórmula de escolhas corretas para cada etapa e tipo de aplicação. A Figura 2.4 apresenta de forma ilustrativa todo o processo KDD.

A seguir uma breve descrição das etapas do processo KDD:

1. Compreensão do domínio: O conhecimento prévio relevante, os objetivos do usuário final e do ambiente em que o processo de descoberta ocorrerá, são essenciais para


FIGURA 2.4 – O processo KDD adaptado de (FAYYAD *et al.*, 1996a)

que se obtenha qualquer informação de valor;

2. Organização do conjunto de dados: Envolve a seleção de um conjunto de dados que serão utilizados para a descoberta do conhecimento, logo inclui saber quais dados estão disponíveis, obter dados adicionais necessários e fazer a integração, incluindo os atributos que serão considerados para o processo;

3. Pré-processamento e limpeza dos dados: Nesta etapa, a confiabilidade dos dados é aprimorada. É um estágio crucial no processo, pois a qualidade dos dados vai determinar a eficiência dos algoritmos de mineração;

4. Transformação dos dados: Após a compreensão, seleção, limpeza e pré-processamento dos dados, são necessários o armazenamento e a formatação adequada para a aplicação dos algoritmos;

5. Função de Mineração de Dados: Nesta etapa se decide qual tipo de algoritmo de mineração de dados utilizar, de acordo com os objetivos do KDD. A principal meta do processo de DM é fornecer informações que possibilitem montar melhores estratégias. A mineração de dados pode desempenhar suas tarefas por meio de dois tipos de modelos:

   - Modelo Preditivo: aprendizado supervisionado, no qual há o uso de variáveis na base de dados para predizer valores futuros;
   - Modelo Descritivo: aprendizado não supervisionado, onde se almeja detectar padrões.

6. O Algoritmo de Mineração de Dados: Após a seleção do tipo de algoritmo, o próximo passo é selecionar o(s) método(s) a ser utilizado (árvore de decisão, agrupamento, redes neurais, dentre outros) e parâmetros apropriados para serem aplicados à procura por padrões nos dados observados;

7. Aplicação do Algoritmo de DM: Nesta etapa do processo há a implementação do algoritmo, que faz a pesquisa para encontrar padrões. Para que se alcance um resultado satisfatório pode ser necessário empregar o algoritmo várias vezes, com ajuste de parâmetros;

8. Interpretação: Após a implementação do algoritmo e a descoberta de padrões, há uma avaliação e interpretação dos padrões, em relação aos objetivos definidos na primeira etapa.

9. Empregando o conhecimento descoberto: Nesta etapa se incorpora o conhecimento ao sistema de desempenho para ações futuras, tomar decisões com base no conhecimento extraído ou documentar às partes interessadas. O sucesso desta etapa determina a eficácia do processo KDD.

# 3 MATERIAL E MÉTODOS

Nesta Seção encontra-se a descrição do processo metodológico utilizado neste trabalho. Representado pela Figura 3.1, este processo se deu pela seleção dos dados, que após o pré-processamento foram representados por uma série temporal, na qual, foi aplicado um método de agrupamento (HMM) para a obtenção de regimes ($R_1, R_2, R_3$). Para geração do modelo de previsão, foram adicionadas variáveis explicativas, previamente selecionadas, à base de dados, as variáveis regime (variável resposta de HMM) e regime anterior (variável obtida a partir da variável regime, em que são obtidos os regimes do dia anterior, isto é, no tempo (t-1). Assim foram aplicados ao novo conjunto de dados classificadores que, após a comparação dos resultados entre estes por meio de técnicas de avaliação dos métodos de classificação, obteve-se o modelo final de previsão.

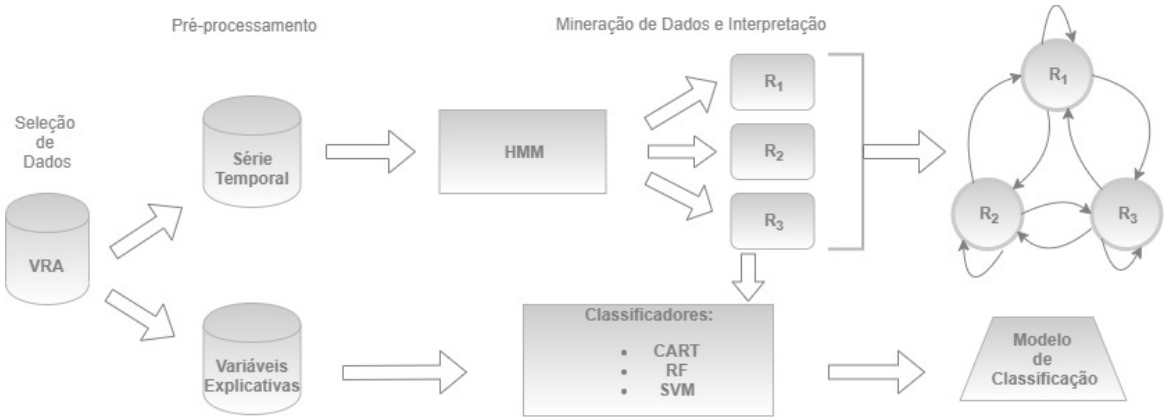



FIGURA 3.1 – Processo metodológico

Para que a exploração dos dados leve à extração de conhecimentos importantes e úteis para a criação de um modelo de previsão da ocorrência de dias congestionados, é necessário um método que possibilite lidar e tratar a base de dados como apresentado na seção 2.6. Neste contexto, o processo KDD foi empregado viabilizando a identificação e associação dos padrões encontrados nos atributos de interesse (conjuntos de determinantes de atrasos e cancelamento de voos), possibilitando previsões em um futuro próximo. O processo KDD foi aplicado empregando a linguagem R, versão 3.5.1, integrada à interface RStudio.

Como sugere a Figura 3.1, o desenvolvimento deste trabalho deu-se em duas fases. Na primeira, foi considerado o conjunto de dados sem o uso de variáveis explicativas, e, de modo não supervisionado, foi empregado o método de HMM para a identificação de estados. Na segunda fase, foram selecionadas variáveis explicativas para a composição da base de dados na qual foram empregados classificadores para a criação do modelo de previsão. A descrição do modelo segue nesta ordem integrado aos passos do KDD.

## 3.1 Primeira Fase

Após o pré-processamento dos dados, um método não supervisionado, de formação de agrupamentos é aplicado, visando a descoberta de padrões.

### 3.1.1 Compreensão do Domínio e Organização dos Dados

Neste de estudo foram considerados, os voos do Aeroporto Internacional de Guarulhos (GRU) que chegam ou partem com mais de 15 minutos do tempo previsto (padrões internacionais de atraso) e voos cancelados. Os voos com horário não programado foram considerados voos realizados sem atraso. O período considerado para a composição da base de dados foram os anos de 2011 a 2017, exceto os meses de junho e julho do ano de 2014, cujos dados não foram fornecidos devido à indisponibilidade de informações de voos regulares e da suspensão do sistema de horário. Tais mudanças no sistema ocorreram em razão do período da Copa do Mundo que se passava no Brasil.

A Agência Nacional de Aviação Civil (ANAC) é uma das agências reguladoras federais do País e foi criada para regular e fiscalizar as atividades da aviação civil. Os dados que compõem a base de dados deste trabalho foram extraídos de relatórios fornecidos pela ANAC e estão disponibilizados a toda sociedade no seu site.

Os relatórios contemplam dados de empresas brasileiras e estrangeiras, com voos de origem ou destino ao Brasil, de voos domésticos e internacionais que foram realizados ou cancelados. Apresentam, entre outras informações, a empresa que operou o voo, o aeroporto de origem, o aeroporto de destino, os horários programados e realizados de pouso e de decolagem dos voos, bem como as causas de eventual atraso ou cancelamento.

Para a composição da base de dados foram selecionados os movimentos diários de chegada e partida do aeroporto, realizados com atraso ou cancelados. Os dados foram organizados, tabulados e, com o uso da linguagem R, foram gerados gráficos para observar como os dados se comportam.

### 3.1.2 Pré-processamento e Limpeza dos dados

É uma etapa crucial em que a qualidade dos dados é analisada, verifica-se a presença de ruídos, inconsistência e ausência de dados. Em um primeiro momento, foi feita a exploração dos dados, e as transformações dos dados necessárias para as análises neste trabalho. Esta etapa foi finalizada com o do tratamento dos valores faltantes e discrepantes (*outliers*).

#### 3.1.2.1 Exploração e Transformação dos Dados

Com o objetivo de observar como os dados originais se comportam foi realizada a análise das variáveis por meio de gráficos de barras comparando as porcentagens de acordo com a variável analisada. Os gráficos contêm informações a respeito da frequência de cada classe, ou porcentagem da variável dependente no eixo vertical.

Na Figura 3.2, observam-se os totais por variável dependente que corresponde ao *status* de voos e pode assumir os valores Atrasado/Cancelado. De acordo com os dados em sua forma original, do total de movimentos de voos programados, 23,39% foram cancelados ou partiram ou chegaram atrasados ao longo dos anos de 2011 a 2017.

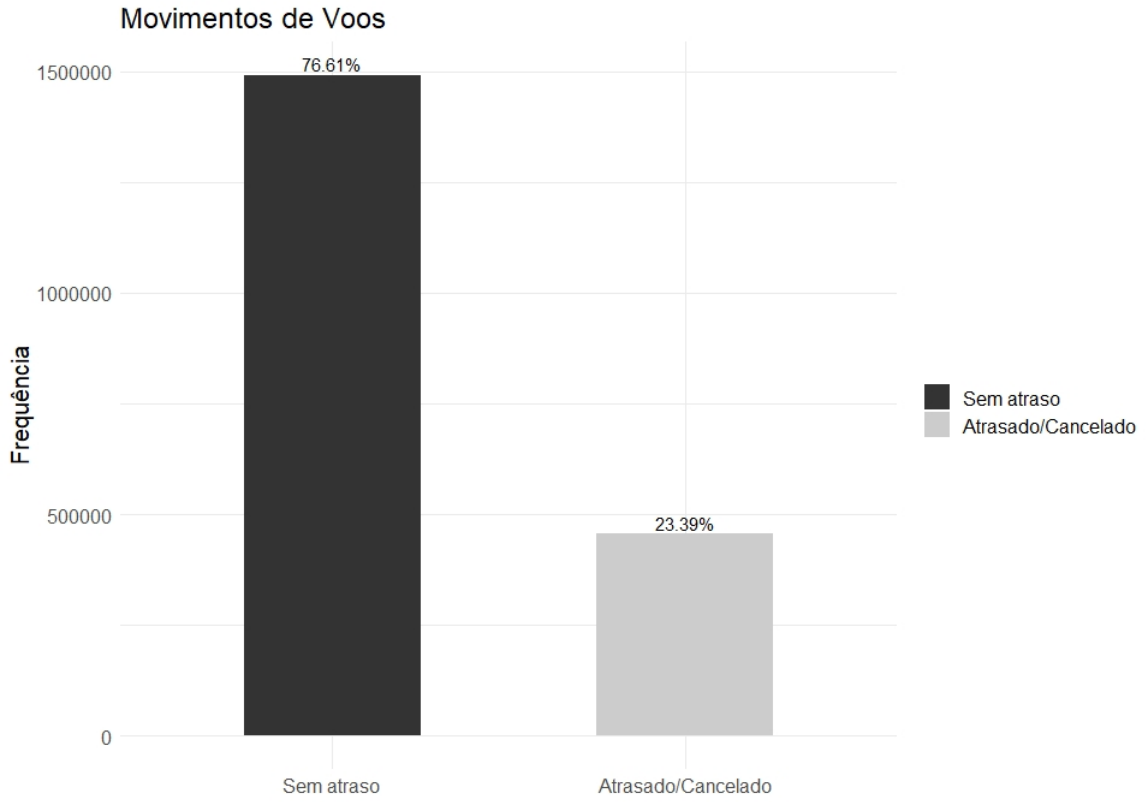



FIGURA 3.2 – Movimento de voos no período de 2011 a 2017, no Aeroporto Internacional de Guarulhos.

Considerando os movimentos de voos anuais e as classes de voos realizados, bem como aqueles realizados com atraso e cancelados, observa-se na Figura 3.3 que os anos de 2011 e 2014 apresentam os maiores índices de atrasos de voos, com 20,17% em 2011 e 17,87% em 2014. No ano de 2011, o Aeroporto Internacional de Guarulhos era administrado pela estatal Infraero (Empresa Brasileira de Infraestrutura Aeroportuária), em fevereiro de 2012 houve a concessão de 51% da administração do aeroporto para a iniciativa privada por 20 anos. Investimentos visando o aumento da capacidade do aeroporto para a Copa do Mundo de 2014 foram feitos e o Terminal 3 foi inaugurado em maio de 2014.

O que diferencia os anos de 2011 e 2014 dos demais, são os fatos de em 2011 o aeroporto ser administrado por uma estatal brasileira e em 2014, já na iniciativa privada, o aeroporto estar passando por reformas e, de modo geral, se preparando para o aumento de demanda em decorrência do evento esportivo mundial. O BRASIL. Departamento de Controle do Espaço Aéreo (DECEA) (2017) e os relatórios dos anos de 2014 a 2018 afirmam que GRU é a principal porta de entrada no país pelo modal aéreo. Ainda, de acordo com as informações apresentadas pelos anuários, o ano de 2014 foi o ano de maior movimento de voos, considerando pousos, decolagens, sobrevoos, toques e arremetidas.

Em relação aos índices de cancelamentos de voos, o ano que apresenta maior taxa é 2015 com 11,56% de cancelamento de voos.

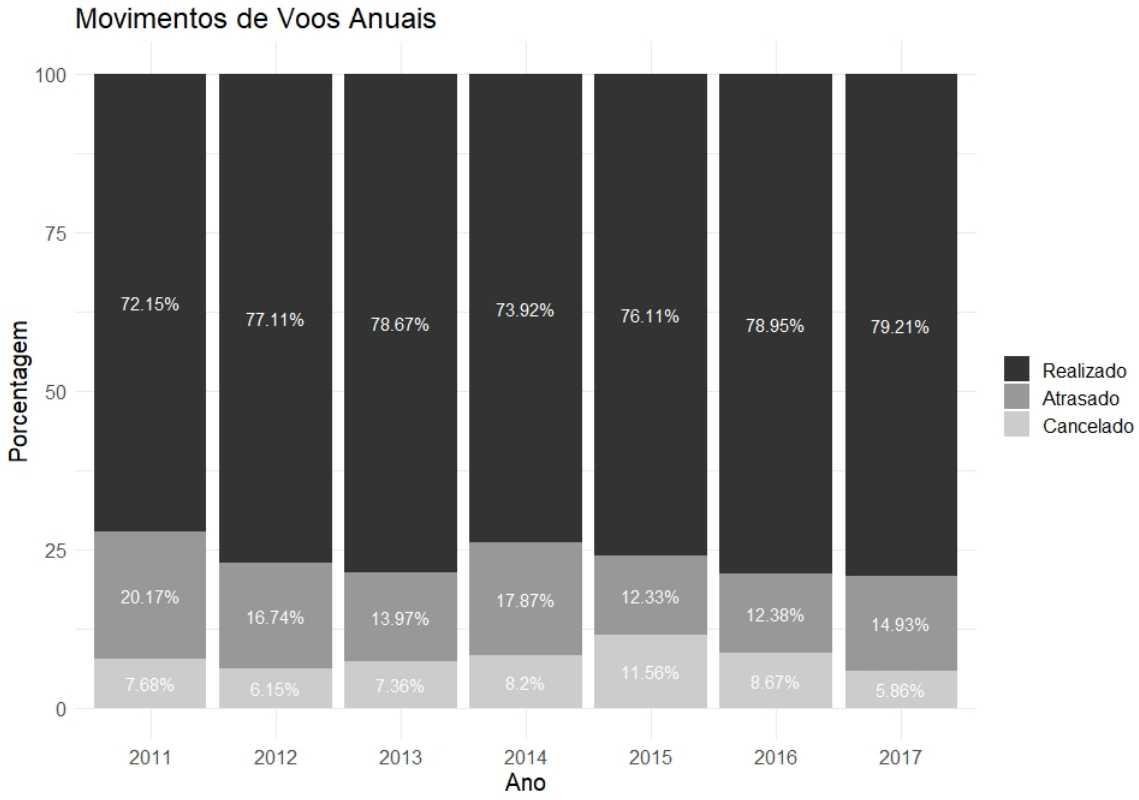



FIGURA 3.3 – Distribuição de voos atrasados e cancelados no período de 2011 a 2017, no Aeroporto Internacional de Guarulhos.

A Figura 3.4 considera o total de voos mensais para os anos de 2011 a 2017. Ao analisar os atrasos de voos a cada mês, observa-se picos de atrasos nos meses de janeiro, julho e dezembro. São meses citados pelos Anuários Estatísticos de Tráfego Aéreo como meses de férias escolares e que normalmente apresentam picos de movimentos de voos.

Rebollo e Balakrishnan (2014) afirmam que o aumento da demanda torna a rede suscetível a atrasos. Os meses que apresentam picos de atrasos de voos, estão em dezembro (20,1%), seguido por janeiro (19,7%) e julho com (16,9%), o que reforça a hipótese de que o aumento de demanda contribui para o aumento de atrasos.

Quanto aos cancelamentos de voos, os meses de abril, junho, fevereiro e maio apresentam os maiores índices com 10%, 9,5% e 9,3%, respectivamente.

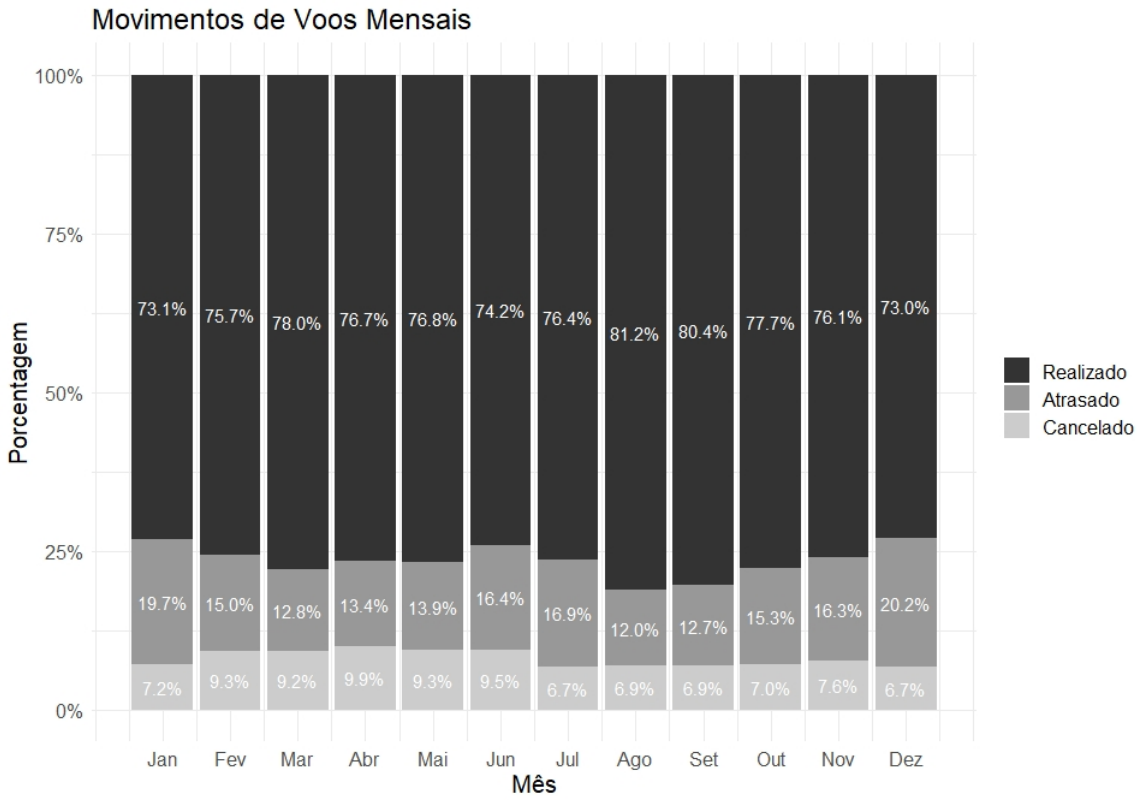



FIGURA 3.4 – Distribuição de voos atrasados e cancelados ao longo dos meses no período de 2011 a 2017, no Aeroporto Internacional de Guarulhos.

Observando os movimentos de voos considerando os dias da semana, na Figura 3.5, nota-se que os dias com maiores porcentagens de atraso são quinta-feira e sexta-feira com 16,9% e 18,1%, respectivamente. De acordo com os Anuários Estatísticos dos anos de 2014 a 2017, os movimentos de voos por dia da semana apresentam os dias de quinta-feira e sexta-feira como os mais movimentados da semana. Ainda de acordo com os anuários, os dias de sábado e domingo apresentam um movimento menor quando comparados aos demais dias. São os dias da semana que apresentam as menores taxas de atrasos de voos, ambos apresentando cerca de 13,5%. Estas observações contribuem com a hipótese de que

a o aumento da demanda contribui para o atraso de voos.

As taxas de cancelamentos de voos estão bastante próximas em todos os dias da semana, apresentando um maior índice nas terças-feiras e quartas-feiras (8,7%) e sábado (8,5%). As menores taxas não se distanciam muito, sendo o menor índice de 7,2% tanto no domingo quanto na segunda-feira.

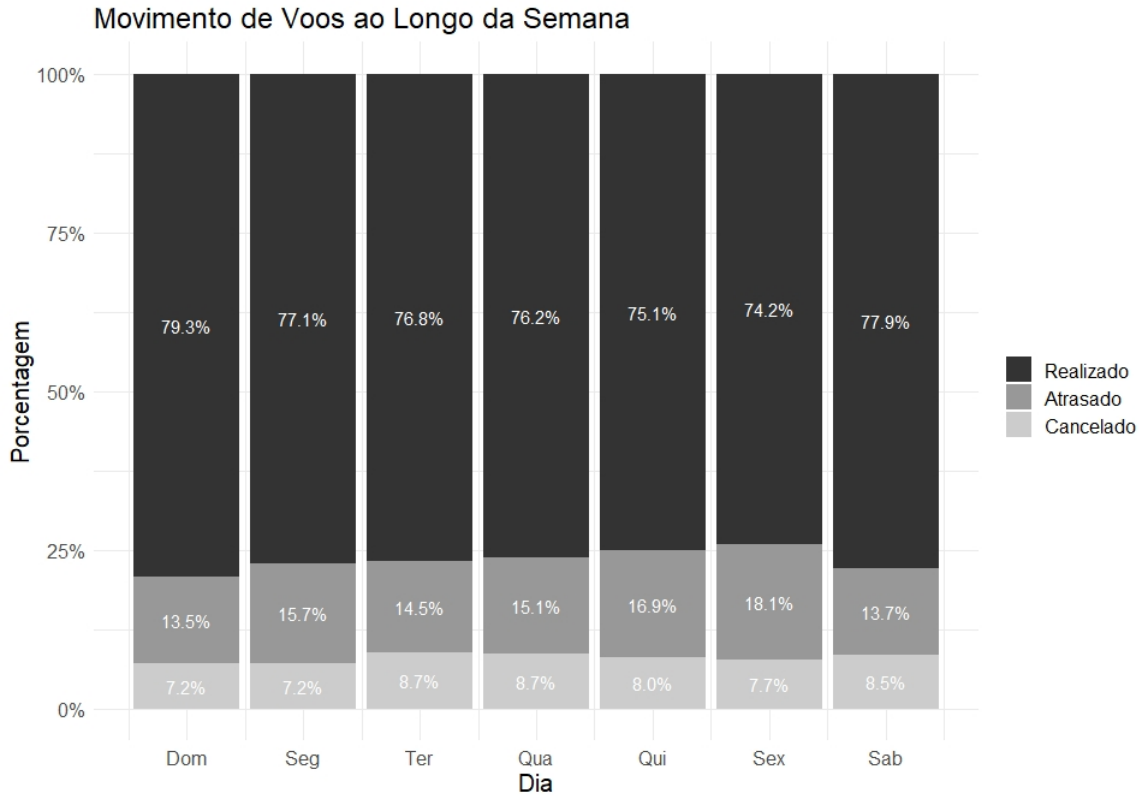



FIGURA 3.5 – Distribuição de voos atrasados e cancelados ao longo dos dias da semana no período de 2011 a 2017, no Aeroporto Internacional de Guarulhos.

A Figura 3.5 apresenta a movimentação de voos no aeroporto GRU nos períodos dos dias úteis e final de semana. Comparando as movimentações, observa-se que durante o final de semana a porcentagem de voos realizados com atraso é menor (13,6%) que nos dias úteis da semana (16,1%). De acordo com o BRASIL. Departamento de Controle do Espaço Aéreo (DECEA) (2017), a movimentação de voos nos dias úteis é maior que no final de semana, assim como oo percentual de voos realizados com atraso. As taxas de voos cancelados durante a semana (8,1%) e final de semana (7,8%) não apresentam uma variação significativa.

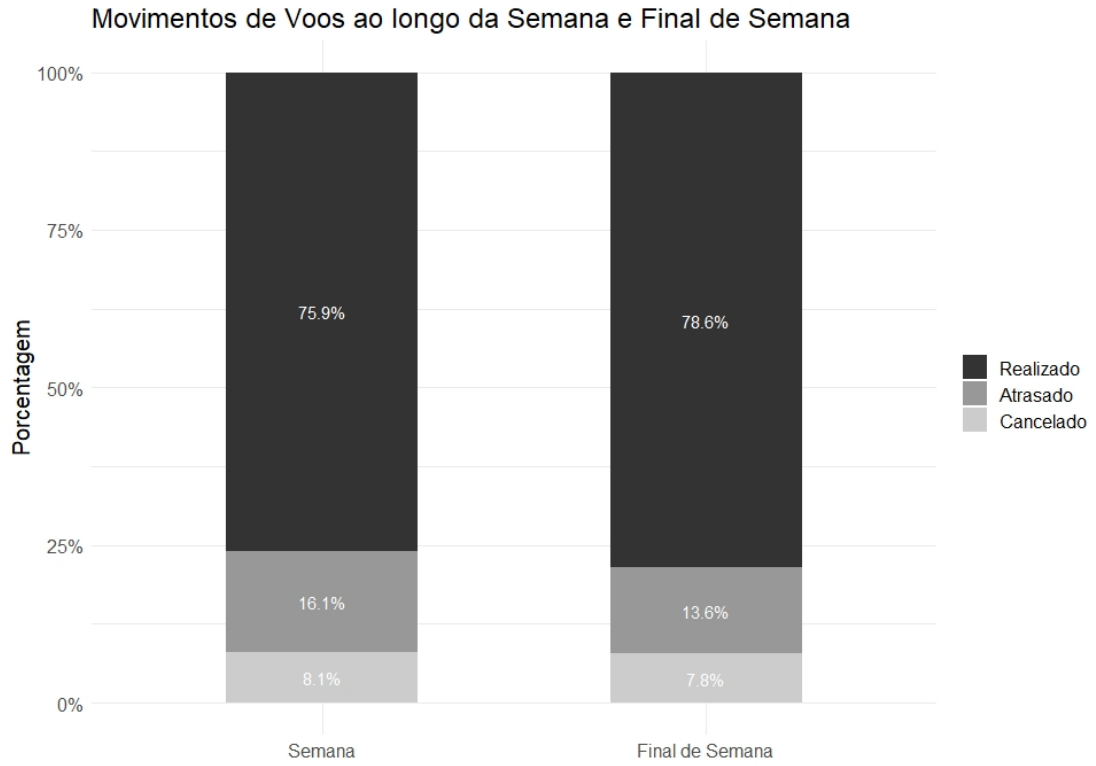



FIGURA 3.6 – Movimento de dias úteis e final de semana no período de 2011 a 2017, no Aeroporto Internacional de Guarulhos.

#### 3.1.2.2 Valores Faltantes

Dados ausentes podem resultar de erro do operador, falhas de sistema de medição, ou do processo de coleta ao longo do tempo, podendo constituir um grave problema para análise. (FAYYAD *et al.*, 1996a).

Para a verificação de observações faltantes do conjunto de dados, foram utilizados os pacotes Mice e VIM, disponíveis na biblioteca do software R. Na Figura 3.7 estão representados os dados das respectivas variáveis de movimentos de voos, realizados, realizados com atraso, cancelados e movimentos programados. Os valores das variáveis estão ordenados de acordo com a variável movimentos programados (Total.geral), em que na escala cinza, quanto mais clara, mais baixos são os valores, e quanto mais escura, mais altos são os valores. Valores faltantes são representados pela cor vermelha, que é observada somente na variável cancelado, não havendo valores faltantes nas demais variáveis.

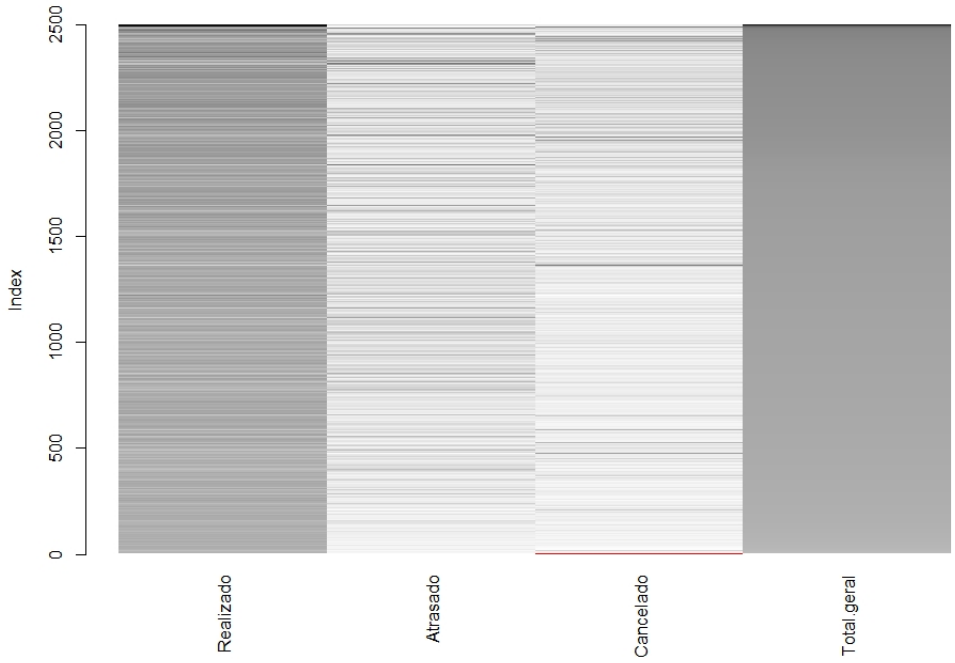



FIGURA 3.7 – Valores faltantes

Para que seja possível determinar o tratamento apropriado aos valores faltantes, é indispensável conhecer a quantidade destes valores para que o conjunto de dados não perca informações que podem ser consideradas de relevância. Na Tabela 3.1 tem-se a representação binária onde o valor 1 indica que não há nenhum dado faltante e o valor 0 representa dado faltante. Verifica-se que das 2496 observações totais da base de dados, 2495 observações não apresentam nenhum dado faltante em nenhuma variável e que apenas uma observação apresenta dado faltante na variável cancelado. A última linha é representada numericamente com o total de dados faltantes em cada variável. Nota-se que há apenas uma observação faltante na variável cancelado.

| *Observações* | *Realizado* | *Atrasado* | *Cancelado* | *Programado* |
|---|---|---|---|---|
| 2495 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 0 | 1 |
| Total | 0 | 0 | 1 | 0 |

TABELA 3.1 – Valores faltantes

Havendo dado faltante em somente uma observação de 2496 observações, há o entendimento de que o número de observações no conjunto de dados é suficiente e não há a necessidade de inclusão de valores. Logo, a observação 1037 (em que consta o dado

faltante) foi excluída do conjunto de dados.

#### 3.1.2.3 Transformação dos dados

Este trabalho almeja gerar um modelo que faça previsões antecipadas de dias congestionados. Entende-se que os transtornos, em função de congestionamentos nos aeroportos, têm como causas principais atrasos e cancelamentos de voos. Logo, são considerados como atributos de interesse do conjunto de dados as variáveis cancelado e realizado com atraso. A variável movimentos programados também é considerada para que seja possível observar a porcentagem de atrasos e cancelamentos em relação aos movimentos programados do dia.

As variáveis são integradas de forma que a série temporal obtida represente a fatia (porcentagem) diária dos voos atrasados e cancelados. A composição dos dados que geraram a série temporal deu-se como segue:

$$fatia = \frac{movimentos \quad atrasados + movimentos \quad cancelados}{movimentos \quad programados}$$

Deste modo, a variável fatia diária de voos atrasados e cancelados, constituída a partir de variáveis de interesse, representa a série temporal univariada que foi trabalhada nesta primeira fase deste estudo.

#### 3.1.2.4 Valores Discrepantes (**Outliers**)

Os *outliers* são observações que numericamente estão distantes do restante do conjunto de dados (valores discrepantes). A detecção destes valores foi feita por meio da análise do *boxplot*. No *boxplot* foram considerados os limites superior e inferior para a determinação de valores que são muito discrepantes (*outliers*) do restante do conjunto de dados. Estes limites são dados por:

1. $LimiteSuperior = Q3 + 1,5 * IIQ$
2. $LimiteInferior = Q1 - 1,5 * IIQ$

Em que,

$Q1 =$ *Primeiro Quartil*;

$Q3 =$ *Terceiro Quartil* e

$IIQ =$ *Intervalo interquartil*, dado por $Q3 - Q1$.

### 3.1.3 Mineração dos Dados

De acordo com Fayyad *et al.* (1996a) a mineração de dados (Data Mining - DM) envolve o ajuste de modelos para determinar padrões a partir dos dados observados e a decisão de que os modelos refletem o conhecimento depende do julgamento subjetivo de um ser humano. Dado o vasto número de algoritmos disponíveis na literatura, visa-se um algoritmo que atenda o modelo de aprendizagem não supervisionado para a detecção de padrões de uma série temporal univariada, que na primeira fase deste trabalho representa a nova base de dados após a transformação.

A Figura 3.8 indica que as observações da série fatia diária de voos atrasados e cancelados são dependentes ao longo do tempo. Zucchini *et al.* (2016) afirmam que, havendo autocorrelação das observações, uma maneira de lidar com a dependência, é relaxar a suposição de que o processo é serialmente independente. Para isso, assume-se que é uma cadeia de Markov, um modo simples e matematicamente conveniente, que resulta em um modelo escondido de Markov.

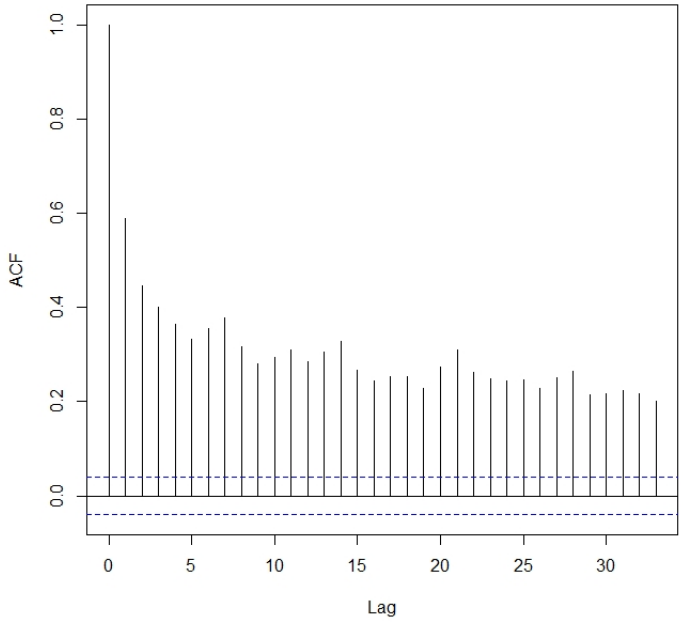



FIGURA 3.8 – Função autocorrelação da fatia de atraso e cancelamento de voos

Os modelos escondidos de Markov são flexíveis para séries temporais univariadas e multivariadas, incluindo séries categóricas e numéricas (ZUCCHINI *et al.*, 2016). Ainda de acordo com o autor, como mencionado na Seção 2.1, este método tem atrativos como simplicidade e fácil tratamento matemático para o cálculo dos parâmetros do modelo. Visser *et al.* (2009. Chap. 13) afirmam que os HMMs tendem a serem aplicados principalmente a séries temporais univariadas longas.

Para que seja possível utilizar HMM, é necessário testar a hipótese de que os possíveis regimes a serem detectados pelo modelo dependem apenas de seu estado anterior, princípio do método de Markov, assim como se há regimes a serem detectados na séria temporal fatia diária de atrasos e cancelamentos de voos que não são observáveis. Sendo assim, aplica-se o HMM para, em seguida, ser testada a hipótese de que o regime atual depende do seu estado anterior.

O modelo HMM foi implementado utilizando-se o pacote *depmixS4*, disponível com maiores detalhes no site *Comprehensive R Archive Network* (*CRAN*). Este pacote incorpora vários tipos de modelos de mistura dependente e permite o ajuste dos HMMs. Nele, é possível escolher entre diferentes famílias de distribuições de probabilidade e como opção padrão a maximização da verossimilhança pode ser feita por meio do algoritmo EM.

#### 3.1.3.1 Algoritmo de Mineração de Dados

O modelo escondido de Markov descreve a evolução de eventos observáveis que dependem de fatores não diretamente observáveis. O método HMM essencialmente avalia a probabilidade de uma sequência de observações, determina a melhor sequência e faz o ajuste de parâmetros de forma que os parâmetros sejam aprendidos sem nenhum conhecimento prévio.

Considerando a série fatia diária de atrasos e cancelamentos de voos, deseja-se encontrar regimes que internamente possuem propriedades em comum que podem representar a intensidade de atrasos e cancelamentos de voos de um determinado período. Para este propósito, utilizou-se o HMM para identificar os possíveis regimes na série temporal. Entretanto, não é possível determinar que regimes foram identificados e nem que sequência foi adotada pelo modelo para se chegar aos regimes obtidos.

#### 3.1.3.2 Critérios de seleção de modelos

Foram utilizados dois critérios de seleção de modelos AIC e BIC para determinar o número de regimes que resulta no melhor modelo para a série de atrasos e cancelamentos de voos. Como mencionado na Seção 2.4.5, o uso dos dois critérios garante a robustez da escolha quando há concordância de ambos em relação ao melhor modelo. A escolha do modelo leva em consideração o ganho de informações condicionado ao aumento de parâmetros, observando que quanto maior o número de parâmetros, mais complexo e custoso é o modelo.

#### 3.1.3.3 O modelo HMM

Neste trabalho, os estados observados foram denominados regimes. Um HMM é composto por estados escondidos, regimes, probabilidades de emissão e transição. A probabilidade de transição indica a probabilidade de ir de um estado para outro ou de permanecer no mesmo estado, e a probabilidade de emissão indica a probabilidade de um regime ocorrer.

A Figura 3.9 é um diagrama simples do modelo HMM, onde foram representadas apenas as probabilidades de transição e os regimes. Considere um modelo com três regimes. As observações diárias de atrasos e cancelamentos de voos podem estar presentes em um dos três regimes, $R_1$, $R_2$ ou $R_3$. Um conjunto de probabilidades de transição está associado a cada regime por meio dos estados e a Matriz de Transição que encontra-se na Seção 2.4.1 apresenta a probabilidade de se passar de um estado $i$ para um estado $j$. Esta formulação gera um modelo que cria uma sequência de regimes, nos quais as observações diárias estão distribuídas.

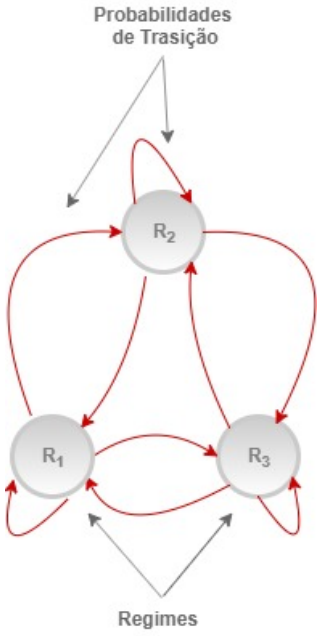



FIGURA 3.9 – Modelo HMM com três regimes

## 3.2 Segunda Fase

Em um segundo momento retorna-se à etapa de pré-processamento do KDD e segue-se para a etapa de mineração de dados. A última etapa explorada por este trabalho, interpretação, é feita por meio da análise dos resultados apresentada no Capítulo 4.

### 3.2.1 Pré-processamento

Ao retomar o pré-processamento, além das variáveis obtidas a partir do modelo de HMM, foi feita a seleção de variáveis explicativas que compuseram a base de dados em que foram aplicados os classificadores.

#### 3.2.1.1 Seleção de Variáveis

Os regimes ($R_1, R_2 e R_3$), detectados pelo modelo de HMM, compõem a variável Regime, esta é a variável resposta desta nova base de dados. A base de dados ainda é composta pelas seguintes variáveis: (i) Regime anterior (que se encontra no instante de tempo $t-1$ e foi extraída a partir variável Regime no instante de tempo $t$ do modelo de HMM); (ii) mês do ano; (iii) dia da semana (que possibilitam investigar se a ocorrência de dias congestionados estão associados a uma demanda maior ou menor durante os dias da semana ou meses do ano); (vi) HHI; (v) Média de spacing; (vi) Desvio Padrão de spacing; e (vii) a Média diária de ConMov. As definições das últimas 4 variáveis são abordadas a seguir.

Rebollo e Balakrishnan (2014) consideram as variáveis dia da semana e mês do ano em sua análise. Ferguson *et al.* (2013) examinou os custos de atrasos de voos em aeroportos observando um dos meses com maior demanda de voos. Abdel-Aty *et al.* (2007) verificou que existem padrões de atrasos semanais e mensais. Os Anuários Estatísticos de Tráfego Aéreo brasileiro dos anos de 2014 a 2018 consideram as variáveis dia da semana e mês do ano em suas análises.

Ainda foram consideradas três variáveis explicativas que, de acordo com Scarpel e Pelicioni (2018), são variáveis potenciais para tratar atrasos e cancelamentos de voos no Aeroporto Internacional de Guarulhos: Índice Herfindal Hirschman (*Herfindal-Hirschman Index* - HHI) de *Slots* por dia, *Spacing*, e *ConMov*. HHI é a variável que mede a concentração de mercado em um aeroporto e refere-se a distribuição da fatia diária de voos operada pelas empresas dentro do aeroporto (SANTOS; ROBIN, 2010).

De acordo com Abdel-Aty *et al.* (2007), *Spacing* é o intervalo de tempo entre dois movimentos de voos programados consecutivos. Neste trabalho, *Spacing* foi considerada como duas variáveis: Média de*Spacing*, variável que representa a média diária de intervalo entre movimentos consecutivos; e o Desvio Padrão de *Spacing*, representando a variabilidade de *Spacing*, ou seja, a diferença no intervalo de tempo entre o movimentos de voo real e o programado. *ConMov* é o número diário de movimentações consecutivas do mesmo tipo (pousos e decolagens) (SCARPEL; PELICIONI, 2018). Neste trabalho foi considerada a média diária de *ConMov*.

Desta forma, o conjunto de dados no qual foram aplicados métodos de classificação é composto pelas variáveis independentes e as variáveis oriundas do modelo escondido de Markov. O conjunto de dados contendo as variáveis e suas definições está representado na Tabela 3.2.

| *Variável* | *Definição* |
|---|---|
| *Regime anterior* | Regime no tempo $t-1$ |
| *HHI* de slots por dia | Índice Herfindal-Hirschman de concentração (concentração de mercado) |
| *Média de Spacing* | Média diária do tempo entre movimentos consecutivos (em minutos) |
| *DesvPad de Spacing* | Desvio padrão diário do tempo médio entre movimentação consecutivas (em minutos) |
| *Média de ConMov* | Número médio diário de movimentações consecutivas do mesmo tipo (pousos ou decolagens) |
| *Mês* | Mês do ano em que o voo está programado (janeiro, fevereiro, março, abril, maio, junho, julho, agosto, setembro, outubro, novembro, dezembro) |
| *Dia da Semana* | Dia da semana em que o voo está programado (domingo, segunda, terça, quarta, quinta, sexta e sábado) |

TABELA 3.2 – Variáveis explicativas do conjunto de dados e suas definições

No modelo de classificação gerado a partir desta base de dados deve-se considerar como uma limitação a incerteza dos rótulos, pois a variável resposta regime e a variável regime anterior não são variáveis concretas e sim variáveis obtidas a partir do modelo de HMM.

### 3.2.2 Mineração de Dados

Os métodos de classificação foram aplicados ao conjunto de dados para as variáveis explicativas apresentadas, em que as classes são dadas pelos regimes identificados pelo HMM. A Figura 3.10 representa o desenvolvimento do DM para a aplicação dos modelos neste estudo.

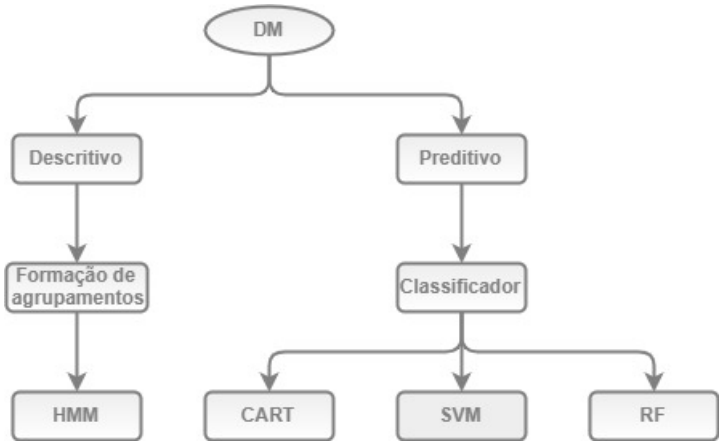



FIGURA 3.10 – Modelos de mineração de dados

#### 3.2.2.1 Métodos de Classificação

Como apresentado na Seção 2.5, os métodos CART, RF e SVM foram escolhidos por serem conceitualmente simples, eficazes para previsões e para a possibilidade de interpretação proporcionada pelo método CART. Os métodos foram aplicados ao conjunto de dados, divido em treino (70%) e validação (30%).

O procedimento de verificação da relevância e seleção das variáveis, e a classificação das observações dos algoritmos CART e RF foram executados simultaneamente. Para a construção do SVM a foram consideradas as variáveis de acordo com a relevância determinados pelos algoritmos CART e RF. Os modelos de previsão por CART e RF e SVM foram implementados utilizando-se os pacotes *Rpart* e *randomForest* e *e*1071.

A fim de se determinar o tamanho ideal da árvore no modelo CART para evitar *overfitting*, foi utilizado o procedimento de validação cruzada 10-*fold* para estimar os erros de previsão. O método considerado para determinar o tamanho ideal da árvore foi a regra "um desvio padrão". O gráfico gerado, considerando este procedimento, é composto pelos erros estimados da validação cruzada versus o parâmetro de complexidade (cp) associado ao tamanho da árvore. O parâmetro de complexidade mede quanta precisão adicional uma partição adiciona à árvore. A precisão é estimada pela combinação linear da taxa de erro e o tamanho da árvore, definida pelo número de nós nos terminais.

De acordo com (MAINDONALD; BRAUN, 2010), o principal hiper-parâmetro do método RF a ser otimizado em modelos gerados pelo pacote *randomForest* é o número *mtry* (número de variáveis testadas aleatoriamente a cada partição), que controla a quantidade de informações em cada árvore individual e a correlação entre elas. O padrão *mtry*, para árvores de classificação, é a raiz quadrada do número de variáveis do modelo. Para otimizar o hiper-parâmetro foram testados diferentes números de *mtry* em função do erro OOB.

O método SVM foi aplicado ao problema de multiclasse utilizando a abordagem um-contra-um. O modelo foi gerado com os parâmetros predefinidos no pacote e utilizando uma função kernel não-linear. Entretanto, foram realizadas otimizações dos parâmetros *custo* e *gamma* almejando um melhor desempenho. Se o valor do parâmetro *gamma* for alto, o raio da área de influência dos vetores de suporte inclui apenas o próprio vetor de suporte e o *custo* não será capaz de prevenir overfitting; se o valor do parâmetro *gamma* for muito baixo, o modelo fica restrito e pode não capturar a complexidade dos dados. Valores elevados para o parâmetro *custo*, podem reduzir os erros de treinamento, entretanto há o aumento do risco de overfitting (JAMES *et al.*, 2013). Para otimizar este parâmetro, foi realizada a validação cruzada onde foram testados valores para os parâmetros *custo* e *gamma*.

#### 3.2.2.2 Interpretação e Avaliação dos Modelos

Após a aplicação dos métodos foi necessário entender o desempenho dos classificadores para o conjunto de dados em estudo. As métricas de erro utilizadas, neste trabalho, foram a matriz de confusão e acurácia. Na matriz de confusão, as linhas representam classes verdadeiras, colunas representam classes preditas, a diagonal da matriz os acertos do classificador e os outros elementos, os erros. A acurácia de um classificador é calculada pela Equação 3.1.

Dado um classificador $l$,

$$acc(l) = 1 - err(l) = \frac{1}{n} \sum_{i=1...n} I(y_i = f(x_i)), \tag{3.1}$$

em que, $n$ é o número de observações,

$I$ a função identidade,

$y_i$ a classe conhecida e

$f(x_i)$ a classe predita.

O método ROC (*Receiving Operating Characteristics*) é uma ferramenta que contribui para avaliar o desempenho dos classificadores. A análise ROC multiclasse (implementada por meio do pacote *pROC*) foi utilizada para comparar o desempenho dos classificadores simultaneamente. A AUC (*Area Under ROC Curve*) é a medida de distância entre as distribuições nas classes a serem avaliadas. Para a análise e visualização gráfica as curvas ROC foram geradas por classe e a área AUC foi calculada pela média das curvas geradas. O gráfico da curva ROC é bidimensional, em que o eixo $x$ representa a Especificidade (taxa de falsos positivos) e o eixo $y$ a Sensibilidade (taxa de verdadeiros positivos). Os

valores produzidos pelas métricas de erro variam entre 0 e 1, em que são considerados melhores, os valores mais próximos a 1.

# 4 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentadas as análises e os resultados obtidos a partir da implementação do HMM, assim como os resultados dos classificadores (CART, RF e SVM) utilizados para gerar o modelo de previsão.

## 4.1 Análise da Série Temporal da Fatia Diária de Voos Atrasados e Cancelados

A Figura 4.1 apresenta a evolução da série temporal da fatia de atrasos e cancelamento de voos entre os anos de 2011 e 2017. No ano de 2014 observa-se nos meses de junho e julho que há dados faltantes. Estes não foram fornecidos pela ANAC em razão da realização da Copa do Mundo no Brasil.

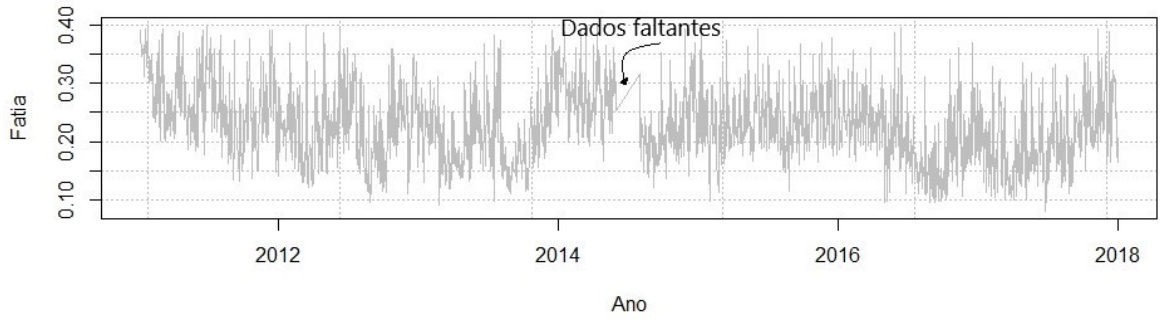



FIGURA 4.1 – Série Temporal da fatia diária de voos atrasados e cancelados para o Aeroporto Internacional de Guarulhos.

No histograma apresentado na Figura 4.2, observa-se que a fatia diária apresenta uma concentração maior de voos em torno de 20%. Há, em uma proporção menor, índices diários que apresentam de 35% a 40% atrasos e cancelamentos.

A Tabela 4.1 apresenta medidas de estatística descritiva da fatia diária de atrasos e cancelamentos de voos. A média da fatia diária, entre os anos de 2011 e 2017, é de

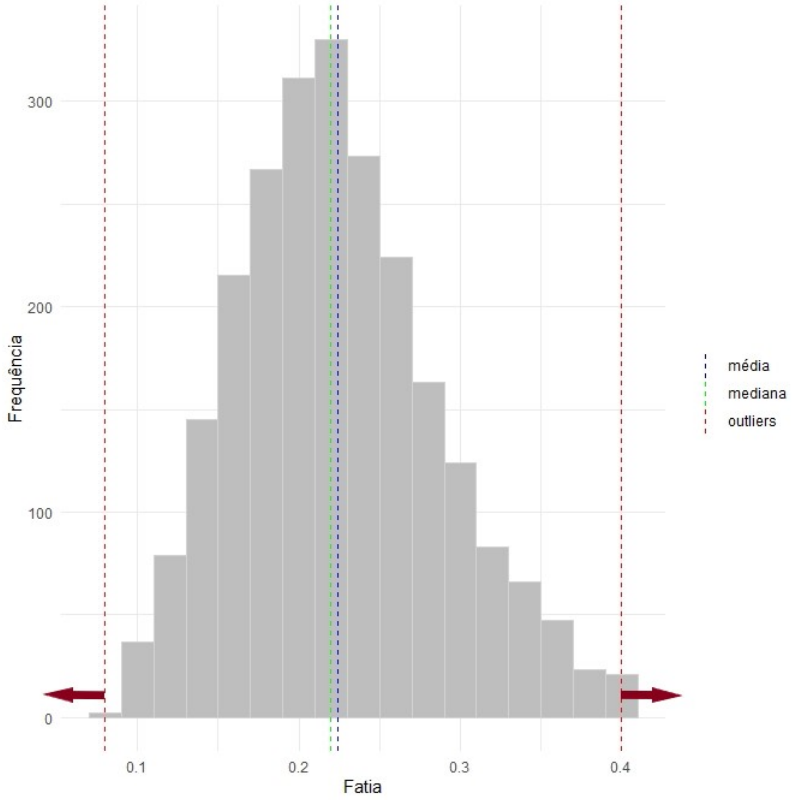



FIGURA 4.2 – Histograma da fatia diária de voos atrasados e cancelados do Aeroporto Internacional de Guarulhos.

aproximadamente 22%. Foram considerados *outliers* os dias que apresentaram um índice acima de 40% da fatia diária de atrasos e cancelamentos e abaixo de 8% (faixa na qual havia apenas uma observação e seu índice era de 4,5%). Tal decisão foi tomada com base na análise do diagrama de *boxplot* da Figura 4.3.

| *Estatística Amostral* | *Valor* |
|---|---|
| Média | 0,224 |
| Mediana | 0,219 |
| Máximo | 0,4 |
| Mínimo | 0,08 |
| Primeiro Quartil | 0,179 |
| Terceiro Quartil | 0,262 |
| Desvio Padrão | 0,062 |

TABELA 4.1 – Estatísticas da fatia diária de atrasos e cancelamentos de voos

Na Figura 4.3, os gráficos foram gerados a partir da variável fatia diária de voos atrasados e cancelados. No gráfico $A$ há a presença de *outliers* e no gráfico $B$ os valores foram removidos. Foram considerados *outliers* as observações da fatia diária que continham

valores acima de 0,4 e abaixo de 0,08. No total foram removidos 85 *outliers* das 2495 observações que compõem o conjunto de dados.

No intervalo entre o terceiro quartil e o limite superior, o índice da fatia diária contém os dias com maior número de atrasos e cancelamentos de voos (aproximadamente entre 23% a 40%). No intervalo entre o primeiro quartil e o limite inferior, o índice da fatia diária contém os dias com menor número de atrasos e cancelamentos de voos (aproximadamente entre 8% a 18%).

Os valores removidos podem sugerir situações interessantes em que o número de atrasos de voos é extremo, entretanto, estes valores são discrepantes do restante do conjunto de dados e podem interferir no resultado do modelo. Para a realização de uma análise a parte, é uma amostra muito pequena, o que geraria um modelo de baixa confiabilidade.

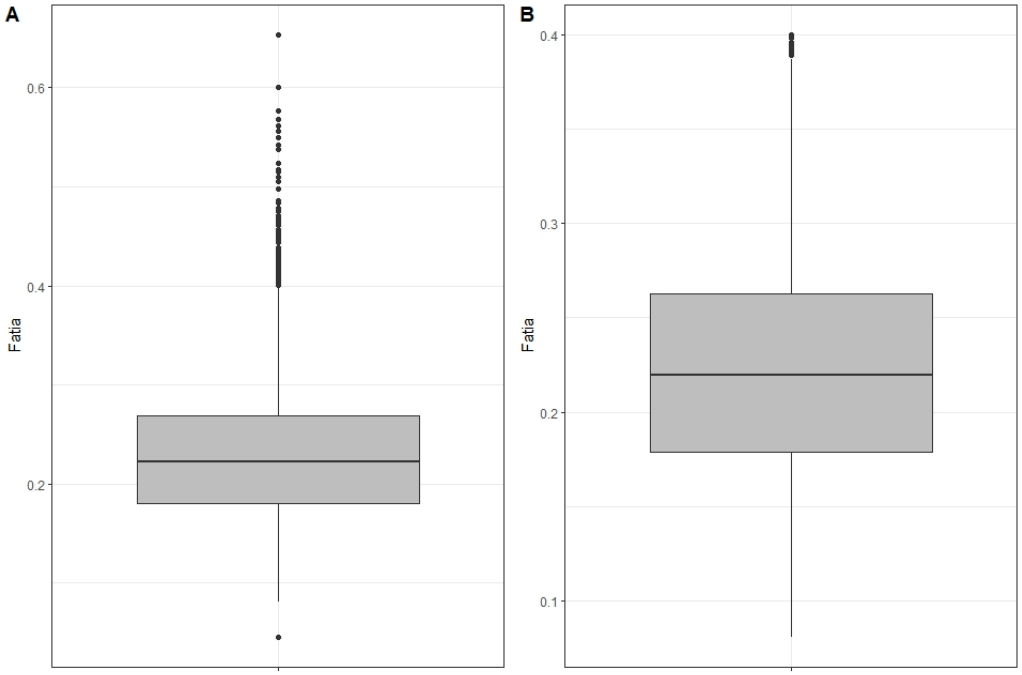



FIGURA 4.3 – Diagrama *boxplot* da fatia diária de voos atrasados e cancelados

## 4.2 Modelo Escondido de Markov

Como mencionado no Capítulo 3, aplicando os modelos escondidos de Markov, decidiu-se testar a hipótese de que os regimes no tempo $t$ dependem de seu estado anterior, ou seja, do seu estado no tempo $t-1$.

### 4.2.1 Ajuste do Modelo HMM

A implementação do modelo depende do número de regimes definidos a *priori*. Para um melhor ajuste, foram avaliados modelos com 2 a 6 regimes, observando os critérios *Akaike Information Criterion* (AIC), *Bayesian Information Criterion* (BIC) e do máximo logaritmo da função de verossimilhança (*Maximum Log-lilelihood - LL*). Um modelo melhor ajustado é aquele com maior LL, entretanto, o aumento do número de parâmetros aumenta a complexidade do modelo. Logo, para ajustar um número adequado de parâmetros, optou-se pela análise dos três critérios simultaneamente.

Os gráficos da Figura 4.4, representam os critérios AIC e BIC que indicam a presençade três regimes. Ao analisar o ganho de informações no gráfico AIC, inferiu-se que a partir de 3 regimes o ganho não é tão significativo quanto do regime 2 para o regime 3. Ao examinar o gráfico BIC, o ganho de informações a partir do regime 3 é ainda menor, assim concluiu-se que 3 é um número de regimes aceitável.

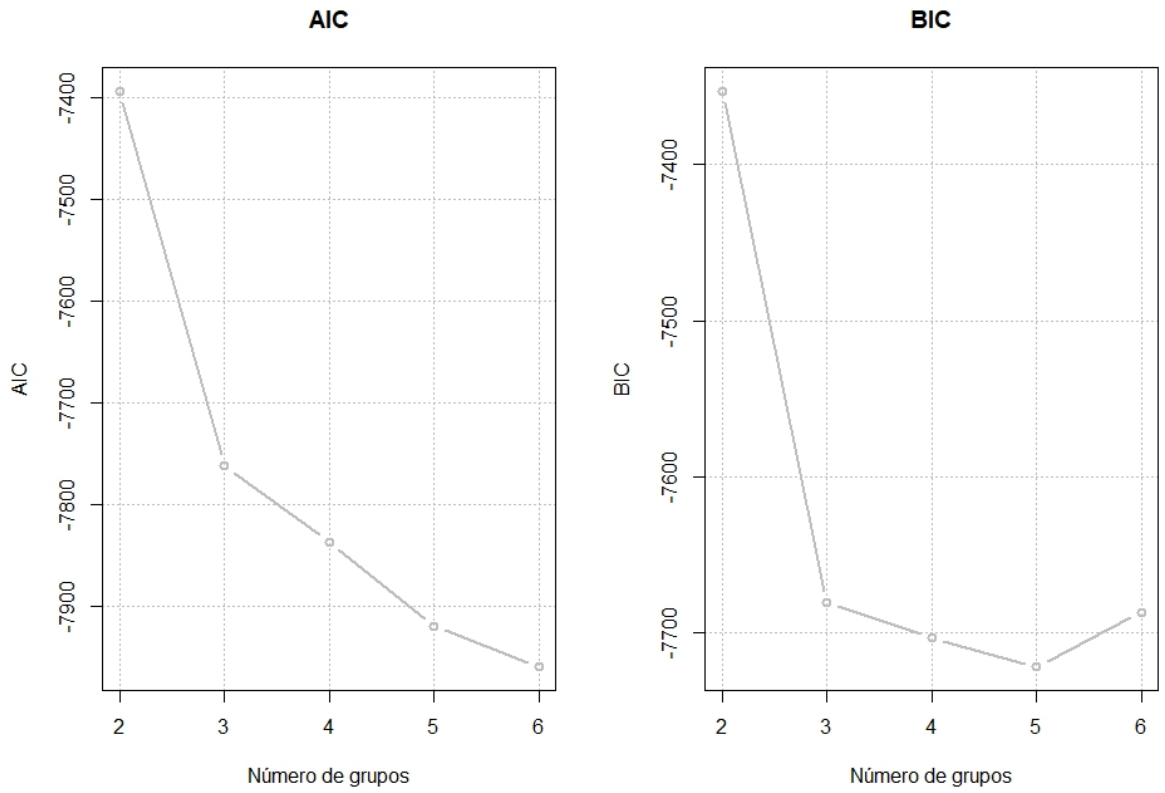



FIGURA 4.4 – Gráfico *Akaike Information Criterion* (AIC) e *Bayesian Information Criterion* (BIC)

Os valores de AIC, BIC, e LL encontram-se resumidos na Tabela 4.2. Quanto maior o número de regimes, maior é o número de parâmetros. A tabela mostra que o maior valor de LL e o menor valor de AIC estão representados por 6 regimes e o menor valor de

BIC está representado por cinco regimes. Entretanto, os modelos com cinco e seis regimes possuem respectivamente 34 e 47 parâmetros.

Considerando os critérios AIC e BIC, nota-se que os valores a partir três regimes decrescem significativamente menos. Pelo critério LL, a partir de três regimes, os valores aumentam sem muita significância, logo, em consonância com os gráficos da Figura 4.4, entende-se que o modelo com 3 regimes e 14 parâmetros é o mais adequado.

| *Regimes* | *Parâmetros* | *AIC* | *BIC* | *LL* |
|---|---|---|---|---|
| 2 | 7 | -7393.112 | -7352.6 | 3703.556 |
| **3** | **14** | **-7761.864** | **-7680.841** | **3894.932** |
| 4 | 23 | -7836.525 | -7703.416 | 3941.263 |
| 5 | 34 | -7918.651 | -7721.88 | 3993.325 |
| 6 | 47 | -7958.964 | -7686.957 | 4026.482 |

TABELA 4.2 – Valores de AIC, BIC, LL e número de parâmetros para os regimes de HMM

Um modelo HMM foi criado para 3 regimes estabelecido *a priori*, de acordo com os critérios de seleção apresentados.

### 4.2.2 Regimes de HMM e Probabilidades Posteriores

A Tabela 4.3 apresenta a média e o desvio padrão de cada regime. Pode-se inferir que o regime 1 é composto pela fatia diária com maior índice de atrasos e cancelamentos de voos e o regime 3 é composto por dias menos congestionados, onde a fatia diária tem o menor índice. Como esperado, o regime 1 apresenta uma variabilidade maior que os demais regimes.

| *Regime* | *Média* | *Desvio Padrão* | *Rótulo* |
|---|---|---|---|
| *1* | 0,297 | 0,046 | Muito congestionado |
| *2* | 0,221 | 0,033 | Congestionamento médio |
| *3* | 0,156 | 0,030 | Pouco congestionado |

TABELA 4.3 – Média e desvio padrão dos regimes de HMM

Para a simplificação da análise, os regimes detectados pelo modelo HMM foram definidos, de acordo com as médias da fatia diária de cada regime, como segue na coluna

"Rótulo"da tabela 4.3.

A Figura 4.5 representa o diagrama de *boxplot* dos regimes identificados pelo modelo de HMM. Os regimes 2 e 3 contêm *outliers* e apresentam menor variabilidade no índice da fatia diária que o regime 3, como visto anteriormente na Tabela 4.3.

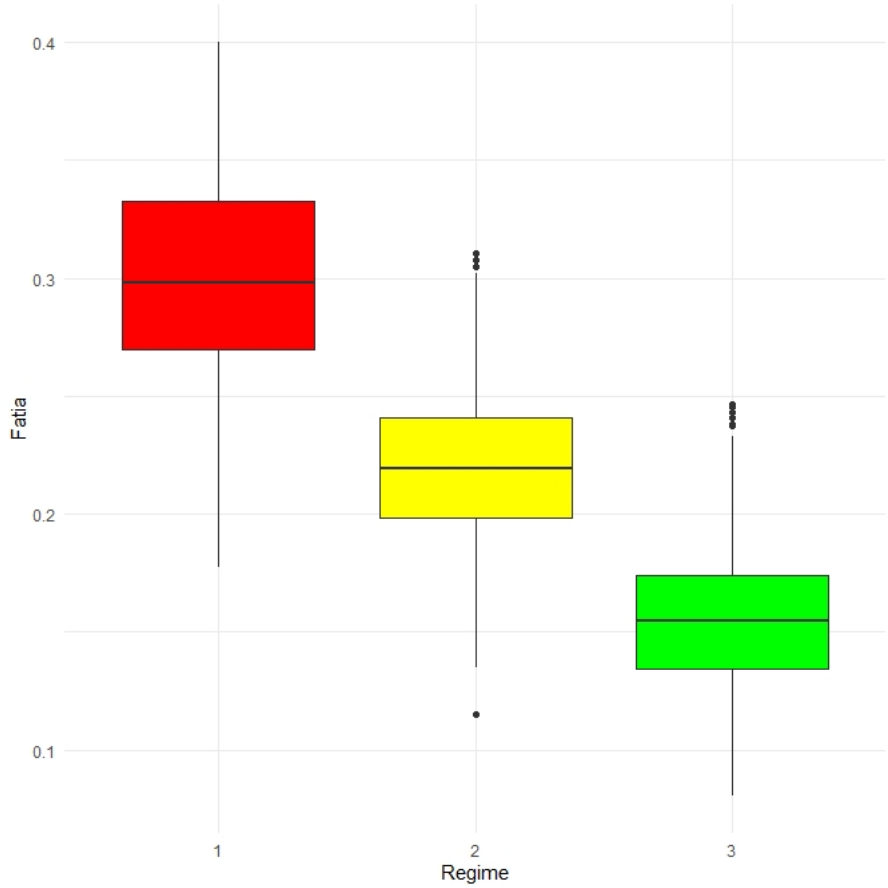



FIGURA 4.5 – Diagrama *boxplot* dos regimes de HMM

A Figura 4.6 apresenta o histograma dos regimes. Como observado na Figura 4.5, o regime 1 apresenta maior variação e maior índice da fatia diária, isto implica a ocorrência de dias muito congestionados neste regime. O regime 2 apresenta a maioria das ocorrências de dias com índices da fatia diária próximos a 22%, o qual indica a ocorrência de dias com congestionamento médio. O regime 3 tem a maioria das ocorrências em torno 15%, não apresentando muita variação dos dias pouco congestionados.

As probabilidades estimadas dos regimes, conhecidas como probabilidades posteriores, são observadas na matriz de transição. A matriz é composta por vetores de probabilidade que são representados pelas linhas, onde cada linha soma um. As probabilidades de transição possibilitam delinear a probabilidade de se estar em um regime particular.

A Tabela 4.4 apresenta a matriz de probabilidades de transição estimadas a partir

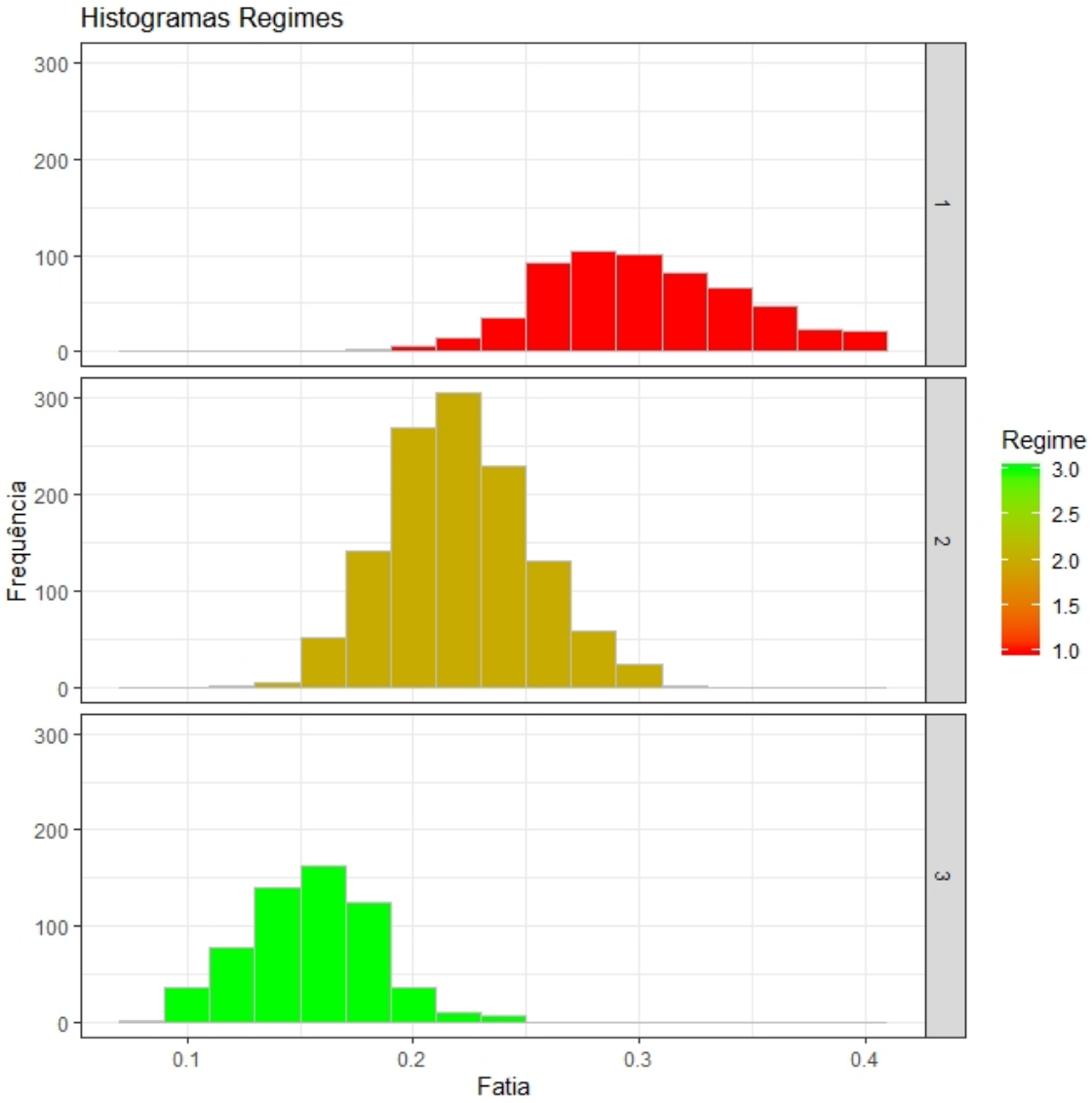



FIGURA 4.6 – Histograma dos regimes de HMM

do modelo. É constituída por três vetores de probabilidade, em que o primeiro vetor indica a probabilidade de um dia muito congestionado permanecer no regime corrente, ou a probabilidade ir deste para o outro regime no instante de tempo $t+1$. O segundo vetor mostra que a probabilidade do regime 2 ocorrer, se este for o regime corrente, é de 80,4%. O terceiro vetor indica que a probabilidade de dias pouco congestionados ocorrer em um próximo instante, sendo 3 o regime corrente, é de 86,6%.

| | | Para | | |
|---|---|---|---|---|
| De | *Regime* | *1* | *2* | *3* |
| | *1* | 0,746 | 0,254 | 0,000 |
| | *2* | 0,127 | 0,804 | 0,069 |
| | *3* | 0,015 | 0,119 | 0,866 |

TABELA 4.4 – Matriz de Transição dos estados de HMM

Nota-se que, de acordo com a matriz de transição, não há a possibilidade de seguir do regime 1 para o regime 3, ou seja, não há chances de um dia pouco congestionado ocorrer em após um dia muito congestionado.

A Figura 4.7 mostra a frequência dos regimes identificados pelo modelo de HMM. O regime 2, com a média diária de 22,1% de atrasos e cancelamentos de voos, tem o maior número de dias do período analisado neste trabalho. O regime 1, que contém os dias muito congestionados, é o menor com uma estreita diferença em relação ao regime 3.

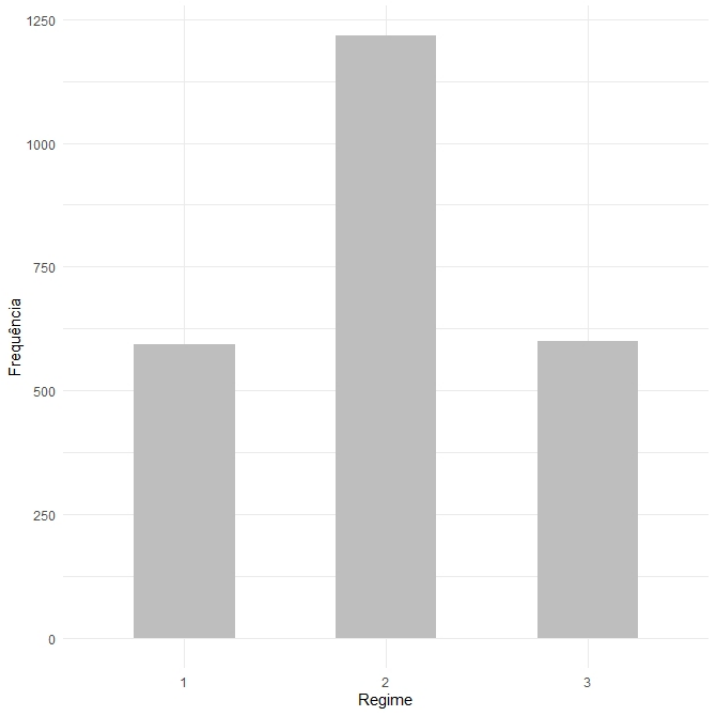



FIGURA 4.7 – Frequência absoluta dos diferentes regimes

A Figura 4.8 apresenta as probabilidades posteriores dos regimes 1, 2 e 3, geradas pelo HMM, sequencialmente ao longo dos anos de 2011 a 2017. É possível observar a predominância dos regimes 2 e 3, e a ocorrência de bastante alternância entre os regimes ao longo do tempo.

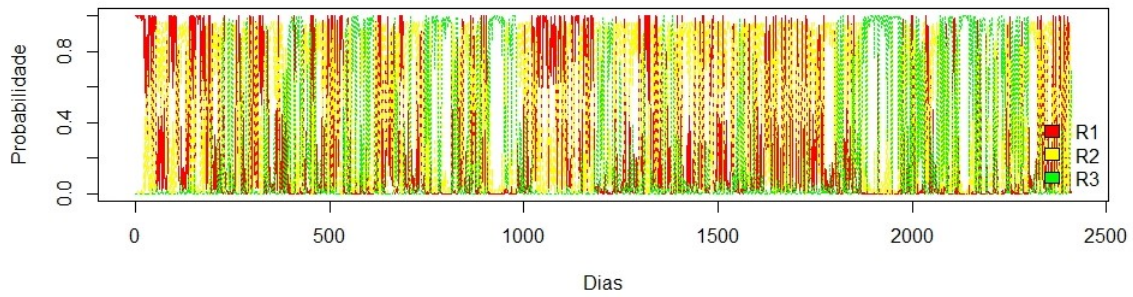



FIGURA 4.8 – Probabilidade posterior dos regimes Série temporal Fatia diária de voos atrasados e cancelados do Aeroporto Internacional de Guarulhos.

Na Figura 4.9 estão representados os primeiros 250 dias do ano de 2011. A escolha dos dias e ano é uma ilustração, apenas para facilitar a comparação visual dos gráficos. Os três gráficos, ao serem comparados, indicam que o modelo identificou bem os três regimes.

No gráfico A está representada a série temporal fatia diária de voos atrasados e cancelados, o gráfico B mostra os regimes detectados pelo modelo e no gráfico C estão as probabilidades posteriores obtidas. O gráfico B é indica que o regime 1 está presente nos primeiros cinquenta dias do ano. Em concordância, o gráfico C mostra que as probabilidades da permanência no regime 1 no mesmo período são altas, próximas a 1. No gráfico A os primeiros cinquenta dias se mostram muito congestionados, pois estão acima de aproximadamente 25% de atrasos e cancelamentos de voos diários.

O regime 3 foi identificado, de acordo com o modelo, como o que apresenta dias menos congestionados. No gráfico B, o regime 3 está presente próximo a 250 dias assim como no gráfico C. O comportamento da série temporal no gráfico A indica que neste mesmo período o índice de atrasos e cancelamentos é menor em relação aos demais dias representados, logo, também está representando o regime 3.

No pacote *DepmixS4* o algoritmo EM considera as probabilidades calculadas a partir dos dados suavizados, as quais são chamadas pelos autores de *smoothed probabilities*. Assim, nos gráficos "Probabilidade Posterior dos Regimes", os regimes não são sempre determinados pela maior probabilidade, pois, são consideradas as *smoothed probabilities*.

Para que se observe o desempenho do modelo ao longo do período analisado, foram gerados gráficos anuais dos regimes identificados pelo modelo e suas probabilidades posteriores. Verificou-se também como os regimes estão distribuídos ao longo de cada ano.

De modo geral, o modelo identificou bem os regimes ao longo dos anos. A probabilidade de que um determinado regime ocorra em um determinado período do tempo é alta e oscila próximo a 1. Na Figura 4.10 nota-se o predomínio dos regimes 1 e 2, indicando que 2011 foi um ano em que dias muito congestionados e dias com congestionamento médio estiveram presentes ao longo de quase todo o ano.

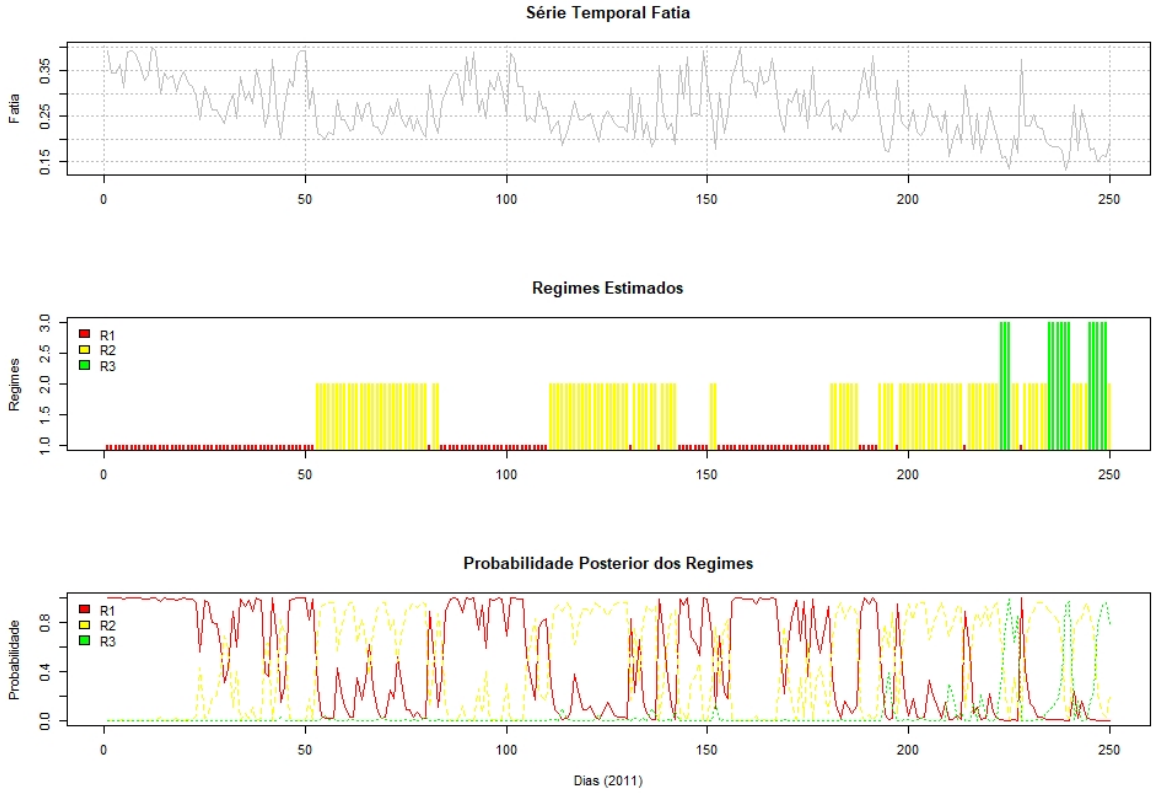



FIGURA 4.9 – Série Temporal Fatia, Regimes Estimados e Probabilidade Posterior dos Regimes dos primeiros 250 dias de 2011.

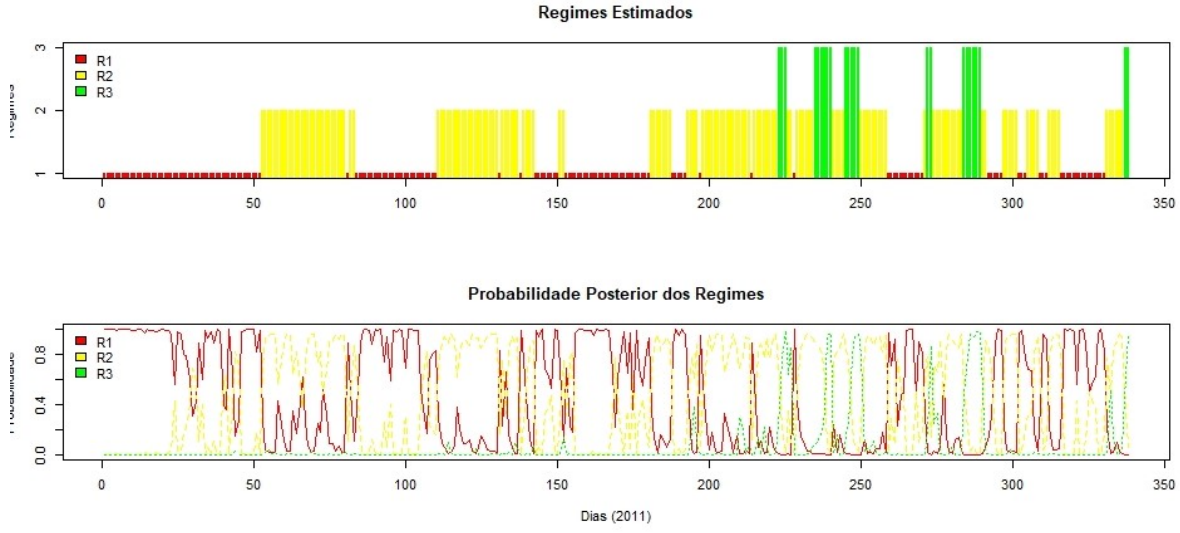



FIGURA 4.10 – Regimes Estimados e Probabilidade Posterior para o ano de 2011.

No ano de 2012, de acordo com a Figura 4.11, houve o predomínio do regime 2 ao longo de quase todo o ano. A ocorrência de dias com o predomínio do regime 1 é menor e há um aumento considerável do regime 3, quando comparado a 2011. Isto é, houve uma diminuição considerável de dias muito congestionados e o aumento na ocorrência de dias

menos congestionados.

A Figura 4.12 mostra que há um aumento gradativo do regime 3, assim como a diminuição do regime 1, se comparado aos os dois anos anteriores. Desta forma, no ano de 2013 houve o aumento da ocorrência de dias menos congestionados e diminuiu a ocorrência de dias mais congestionados.

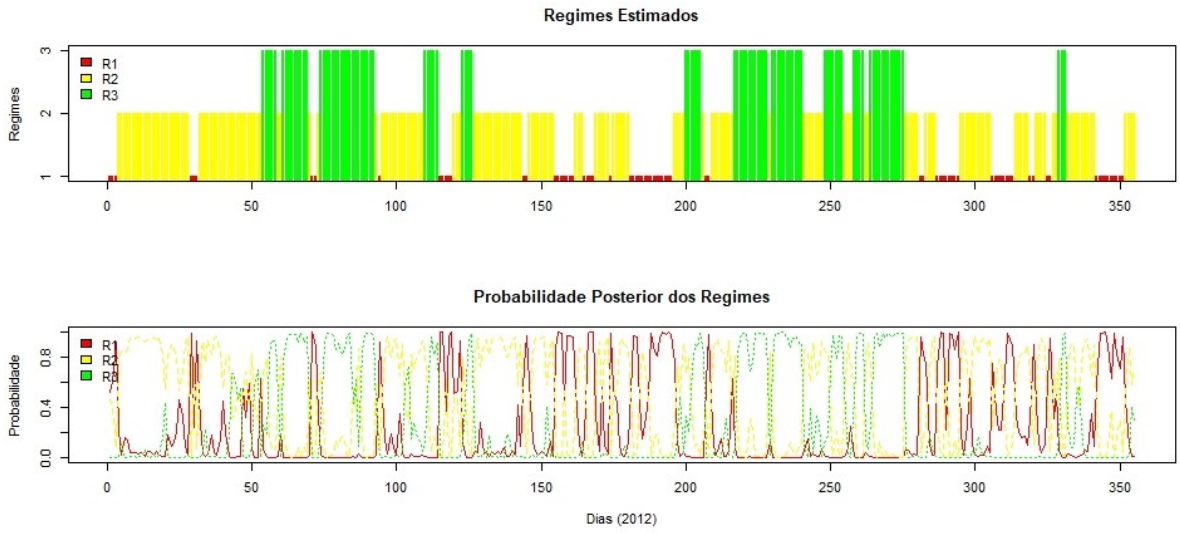



FIGURA 4.11 – Regimes Estimados e Probabilidade Posterior para o ano de 2012

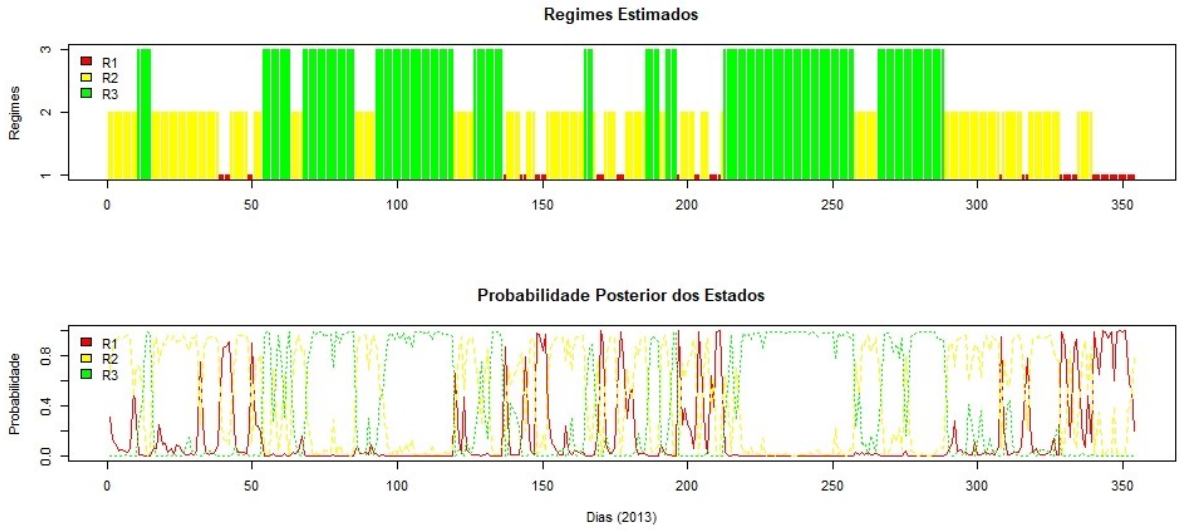



FIGURA 4.12 – Regimes Estimados e Probabilidade Posterior para o ano de 2013

No ano de 2014 houve um aumento significativo de regime 1 e o regime 3 esteve pouco presente, assim como no ano de 2011. A Figura 4.13 mostra que houve o predomínio do regime 1 nos primeiros 150 dias do ano e do regime 2 ao longo do período restante. Logo, houve a ocorrência de dias muito congestionados e com congestionamento médio ao longo de quase todo o ano.

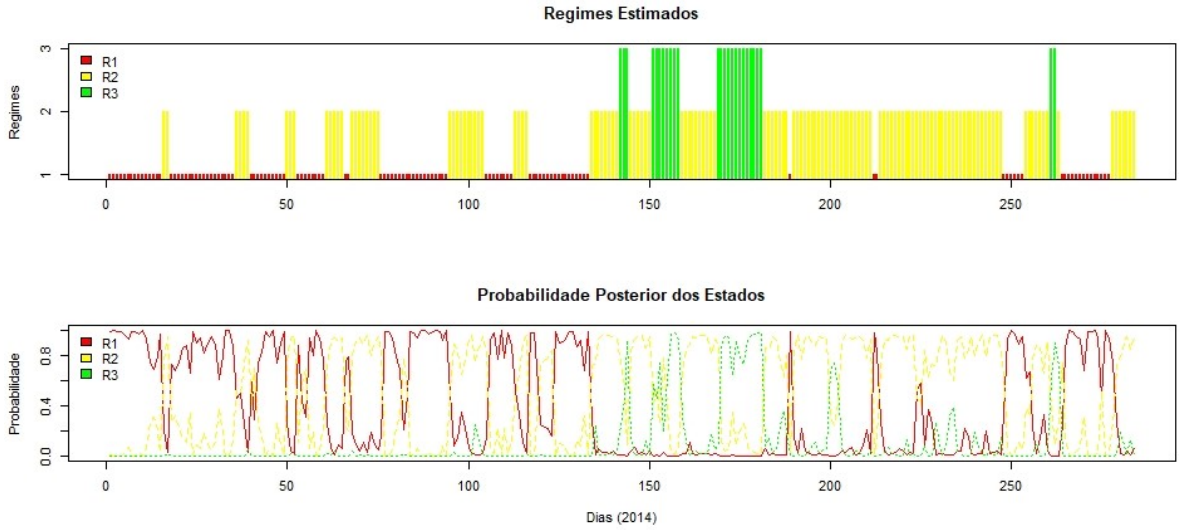



FIGURA 4.13 – Regimes Estimados e Probabilidade Posterior para o ano de 2014

No ano de 2015 houve o predomínio do regime 2 e uma diminuição considerável do regime 1. A Figura 4.14 indica que foi um ano em que houve o predomínio de dias com congestionamento médio e a presença de dias muito congestionados ao longo de todo o ano.

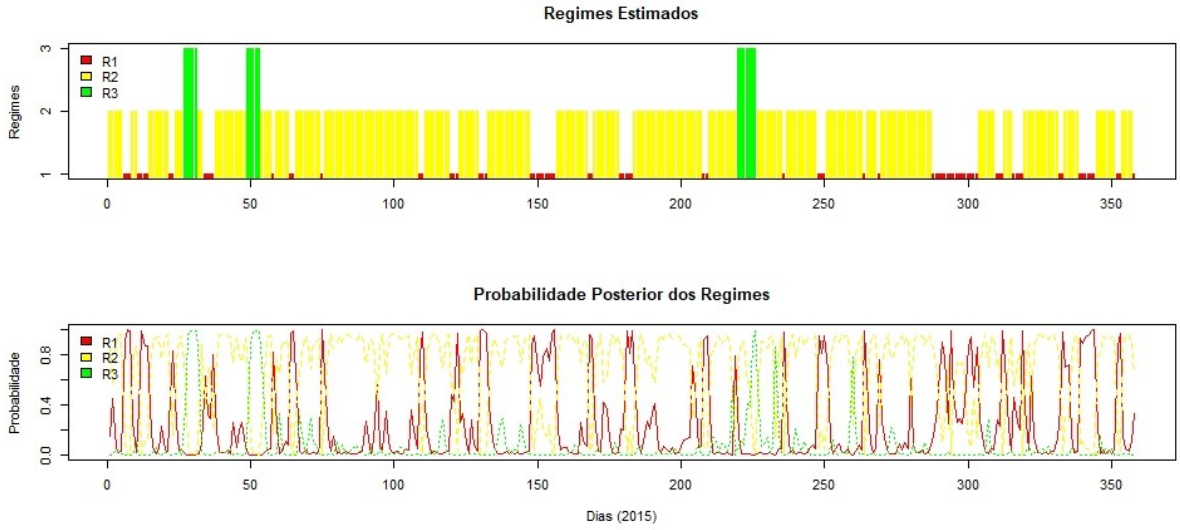



FIGURA 4.14 – Regimes Estimados e Probabilidade Posterior para o ano de 2015

No ano de 2016 o regime 3 aumentou consideravelmente, a Figura 4.15 mostra que houve o predomínio dos regimes 3 entre 200 e 300 dias do ano e do regime 2 ao longo do restante do tempo. Assim, foi um ano no qual prevaleceram os dias com congestionamento médio e pouco congestionados, transcorrendo entre os índices dos regimes 2 e 3.

A Figura 4.16 mostra que o ano de 2017 é similar ao ano de 2016, exceto ao fato de que no ano de 2017 o regime 3 predominou nos primeiros 250 dias do ano. Assim como

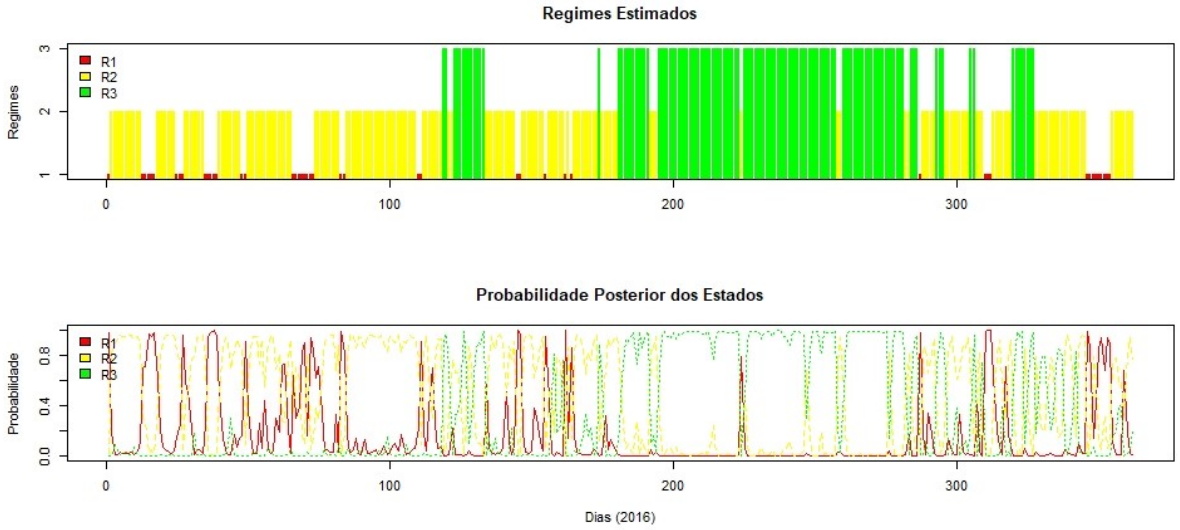



FIGURA 4.15 – Regimes Estimados e Probabilidade Posterior para o ano de 2016

2016, foi um ano em que os índices de atrasos e cancelamentos predominantes oscilaram entre os regimes 2 e 3.

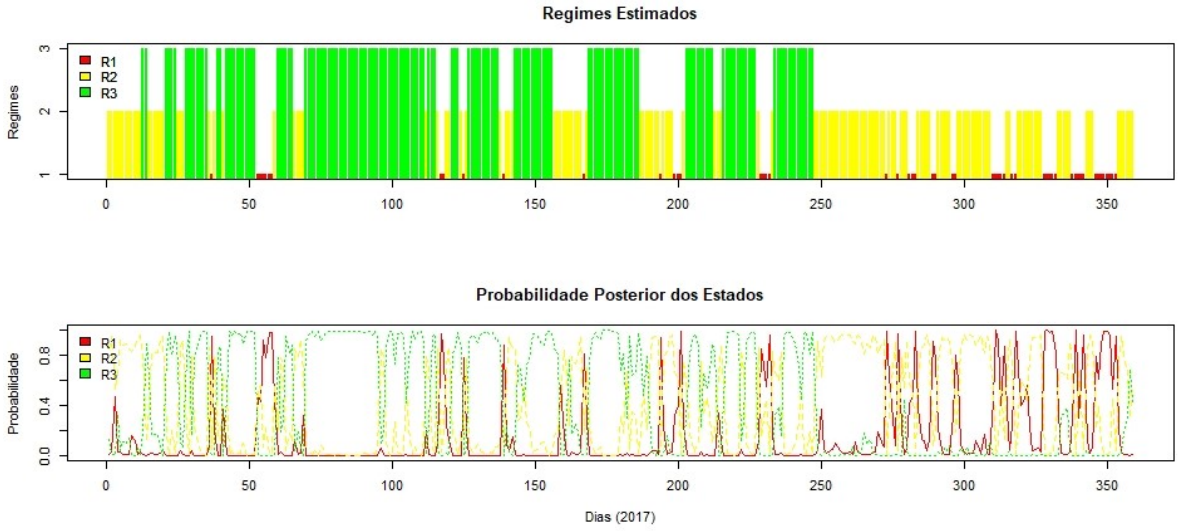



FIGURA 4.16 – Regimes Estimados e Probabilidade Posterior para o ano de 2017

Os gráficos que identificaram os regimes ao longo dos anos mostram que não é possível identificar padrões de distribuição dos regimes anualmente. O ano de 2011 apresentou dias muito congestionados, que foram diminuindo gradativamente ao longo doa anos de 2012 e 2013. Entretanto, o ano de 2014 não deu continuidade ao comportamento dos dois anos anteriores, apresentou ao longo de quase todo o período dias muito congestionados. O ano de 2015 ainda seguiu com a ocorrência de dias muito congestionados e com congestionamento médio, porém com índices predominantemente do regime 2. Os anos de 2016 e 2017 retomaram o aumento da ocorrência do regime 3, todavia de modo geral, em períodos distintos. Cada ano exibiu uma distribuição particular dos regimes.

A Figura 4.17 mostra a distribuição dos regimes ao longo dos meses dos anos de 2011 a 2017. Os gráficos indicam ao longo dos anos, que nos meses de janeiro, novembro e dezembro há o predomínio da ocorrência de dias muito congestionados e com congestionamento médio, onde predominam os regimes 1 e 2. Dias pouco congestionados (regime 3) estão presentes nos meses de agosto, setembro e outubro ao longo de quase todos os anos.

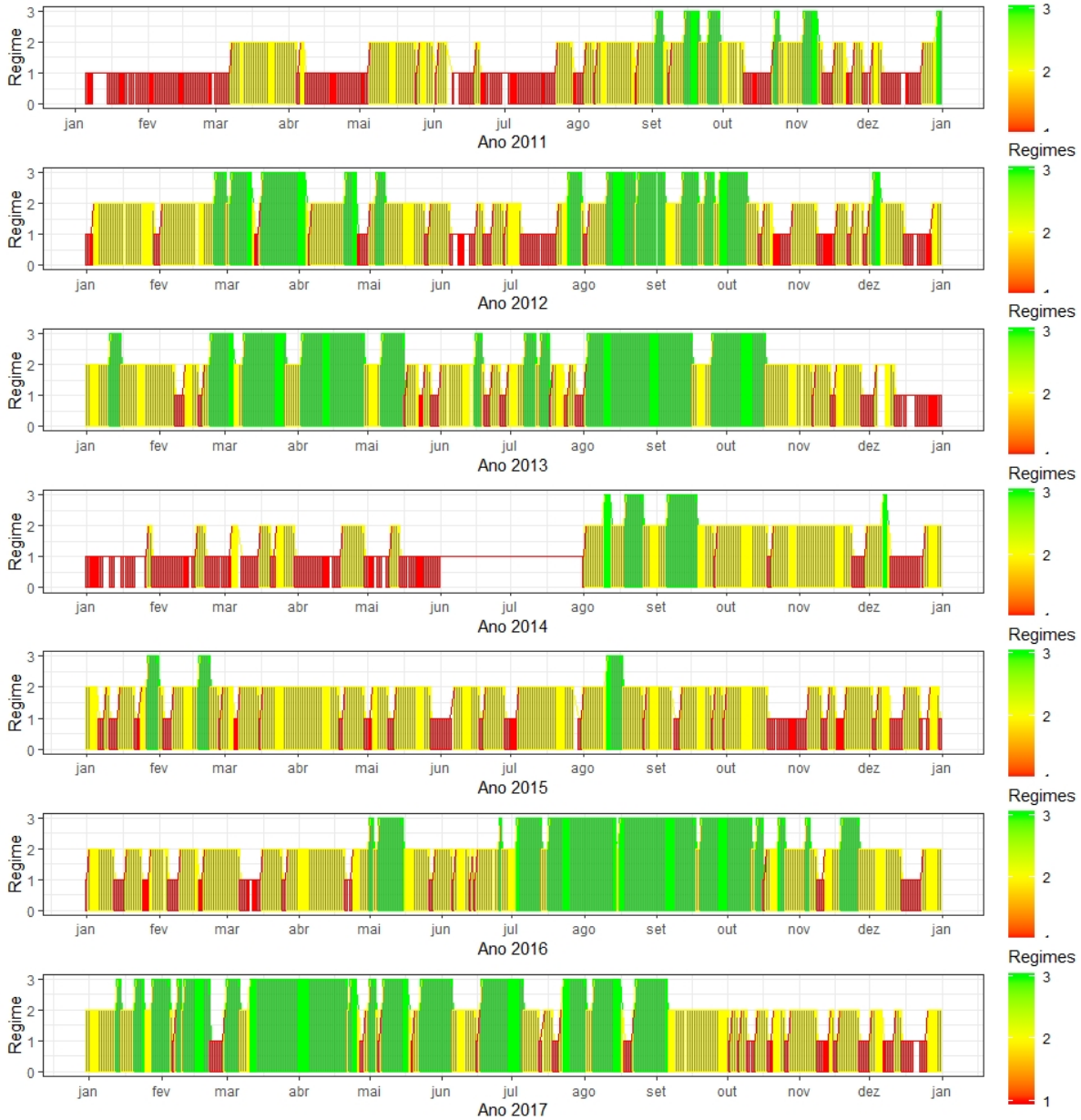
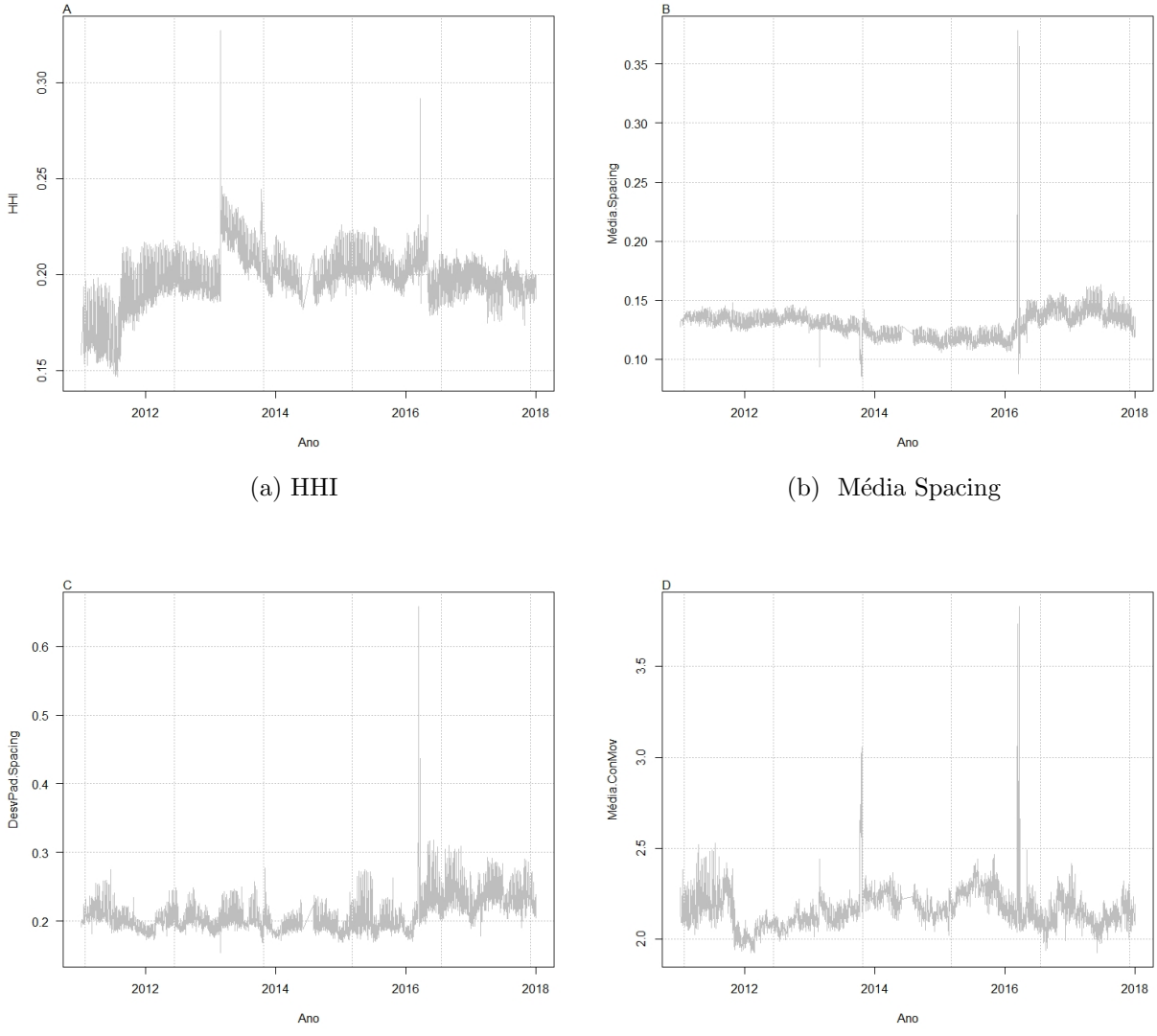



FIGURA 4.17 – Distribuição dos regimes estimados ao longo dos meses.

## 4.3 Modelo de Previsão

Após a aplicação do HMM foi criado um modelo de classificação em que a variável resposta são os regimes ($R_1, R_2 e R_3$) detectados pelo modelo HMM. Portanto, o objetivo deste modelo é prever qual será o regime no instante de tempo $t + 1$.

A Figura 4.18 apresenta a evolução temporal das variáveis independentes contínuas, consideradas neste trabalho: HHI - Figura (a), média diária *spacing* - Figura (b), desvio padrão *spacing* - Figura (c) e média diária de movimentos consecutivos do mesmo tipo - Figura (d). As variáveis, média diária *spacing* e desvio padrão *spacing*, foram multiplicadas por 100 devido ao padrão de arredondamento do pacote do R que gerou o modelo CART.




(a) HHI

(b) Média Spacing

(c) Desvio Padrão Spacing

(d) Média ConMov

FIGURA 4.18 – Evolução temporal das variáveis independentes

### 4.3.1 Árvores de Classificação e Regressão

Na Figura 4.19, a linha horizontal pontilhada indica o valor onde o nível do erro é atingido considerando a regra, o que indica que a árvore deve conter seis nós terminais.

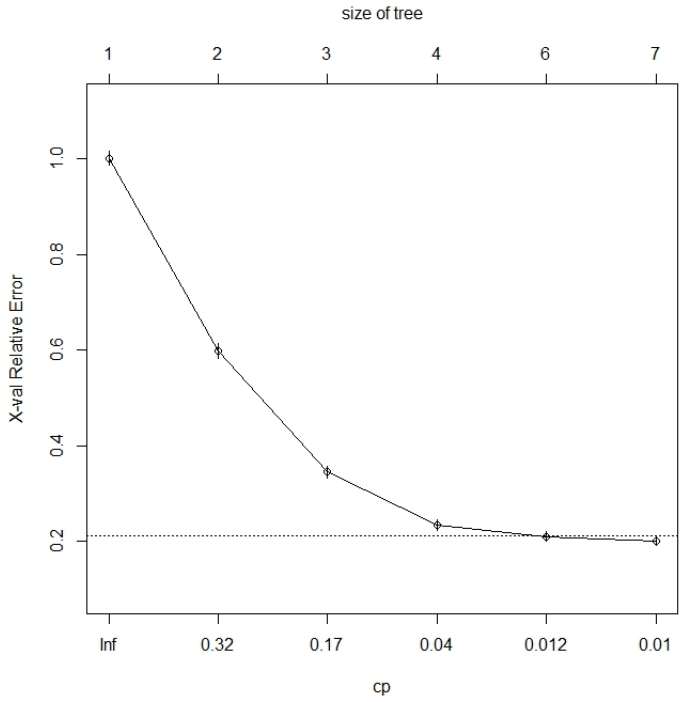



FIGURA 4.19 – Validação cruzada versus parâmetro de complexidade

A Figura 4.20 apresenta a árvore após a realização da poda, com seis nós terminais e cinco partições. É possível observar a indicação de que a variável regime anterior (período de tempo t) tem maior relevância e portanto forte influência na previsão de regimes no instante de tempo $t + 1$. O nó terminal 3, onde há 698 observações, foi classificado como regime 1 e depende somente da variável regime anterior. Desta forma, o dia que pertence ao regime 1 (com a maior média de atrasos e cancelamentos, aproximadamente 30%), tem a probabilidade de 88,1% de que o dia seguinte permaneça congestionado e pertença ao mesmo regime. A taxa de erro é de 11,9%.

O nó terminal 11 (com 712 observações), classificado como regime 3, depende somente de que o regime prévio seja o regime 3. Em um dia, considerado mais tranquilo, com a menor taxa média de atrasos e cancelamentos (aproximadamente 15%), a probabilidade de que o dia seguinte pertença ao regime 3 é de 94%, e a taxa de erro é de 6%. Observe que não há a probabilidade do regime 3 ir para o regime 1, logo, há a possibilidade de um dia muito congestionado preceder a um dia considerado mais tranquilo é 0.

Para que se interprete o nó terminal 10 de forma clara, são abordados brevemente os conceitos de coordenação de *slots* e estrutura de *bank* relacionados ao funcionamento dos aeroportos. A coordenação de *slots* é uma política de gerenciamento dos aeroportos,

que estão no seu limite de capacidade, com o objetivo de maximizar a utilização da estrutura aeroportuária disponível e garantir mais eficiência nas operações do aeroporto (INTERNATIONAL AIR TRANSPORT ASSOCIATION, 2017).

O Aeroporto Internacional de Guarulhos está classificado como nível 2 (aeroporto facilitado), no qual cooperação e alterações voluntárias na programação são necessárias para evitar congestionamento (SCARPEL; PELICIONI, 2018). A estrutura de *bank* é a concentração de voos de chegada ou partida, no mesmo *slot*, em um curto período de tempo para permitir maior conectividade entre os voos (ATER, 2012).

O nó terminal 10 apresenta o maior número de observações (846), é o único nó terminal com alta probabilidade (82,4%) de pertencer ao regime 2. Ocorre quando o regime prévio é 2 e a demanda de movimentos de voos programados é maior (Média *Spacing* $<$ 0,154). Quando a demanda é maior os intervalos de tempo entre os movimentos de voos (pouso e decolagem) tendem a ser menores. Em casos de aumento de demanda, uma alternativa para gerenciar os atrasos é o gerenciamento de *slots* para melhorar a eficiência das operações. Em caso de uso da estrutura *bank* na escolha dos horários no aeroporto, a companhia aérea deve considerar a capacidade da infraestrutura, evitando atrasos nos horários de pico.

O nó terminal 6 (93 observações) foi classificado como regime 1, isto ocorre quando o regime anterior é 2 e a demanda de movimentos de voos diários programados é menor (Média *Spacing* $\geq$ 0,154) nos dias de domingo e sexta-feira. Deste modo, com o intervalo de tempo entre os movimentos diários programados maior do que dois minutos, o dia pertencente ao regime 2 tem 98,9% de probabilidade de ocorrer no regime 1. Segundo Scarpel e Pelicioni (2018), os resultados de seus estudos sugerem que nos dias domingo e sexta-feira também são esperadas maiores taxas de atrasos.

No nó terminal 8, onde há 139 observações, a probabilidade de que o regime 1 ocorra é de (61,2%). Há esta probabilidade de ocorrer o regime 1 quando o regime anterior é 2, a demanda do movimento diário de voos programados é menor (Média *Spacing* $\geq$ 0,154) nos dias segunda-feira a quinta-feira e sábado, e o mercado menos concentrado (HHI $<$ 0,21). Estes resultados estão de acordo com a literatura, pois em dias com uma demanda maior (períodos de pico) e o mercado menos concentrado (mais companhias aéreas operando), é esperado que ocorram mais congestionamentos.

O último nó terminal é 9, com 70 observações, têm probabilidade de 84,3% de pertencer ao regime 3, com erro de 15,7%. O que o diferencia do nó terminal 8 é o mercado mais concentrado (HHI $\geq$ 0,21), no qual as companhias aéreas tendem a internalizar os atrasos. De acordo com (SCARPEL; PELICIONI, 2018) é esperado que taxas menores de atrasos ocorram quando o aeroporto é mais concentrado e a demanda é menor, logo, os resultados estão de acordo com a literatura, pois o nó terminal 9 tem a menor taxa média da fatia

diária, o mercado mais concentrado e a demanda menor. Segundo os autores, dias como estes, atribuídos ao nó terminal 9 neste caso, podem ser nomeados como dias regulares de baixa movimentação, com a maioria dos voos realizados pelas três maiores companhias aéreas brasileiras (Azul, Gol e TAM).

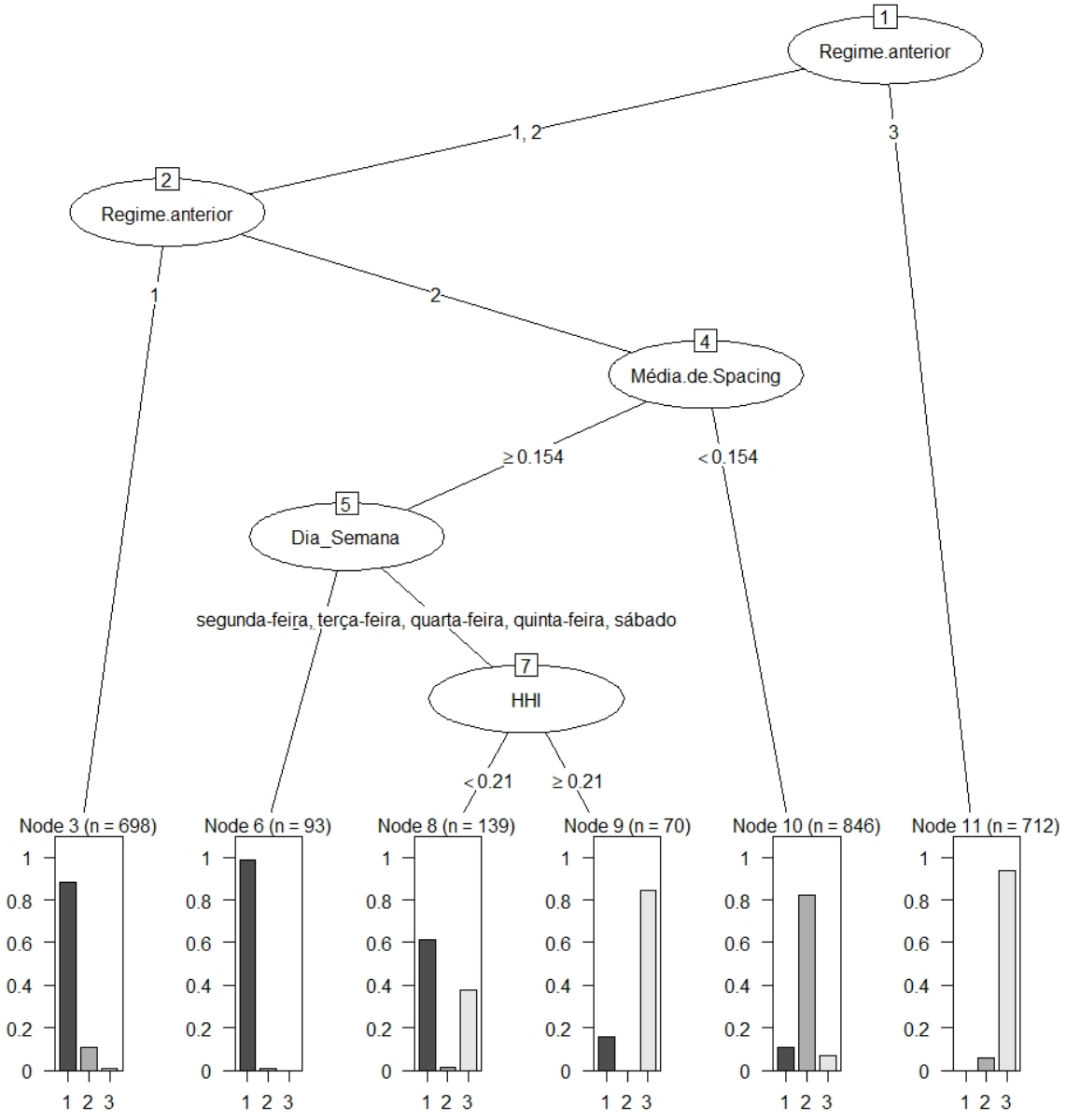



FIGURA 4.20 – Árvore de classificação com seis nós terminais

A Tabela 4.5 apresenta as matrizes de confusão de treino e teste obtidas a partir do modelo CART. Em comparação, as matrizes de treino e teste apresentaram, por classe, respectivamente os seguintes erros percentuais: (i) regime 1: 11,4% e 10,9%; (ii) regime 2: 14,9% e 13,1%; e (iii) regime 3: 13,9% e 10,9%. A acurácia da matriz de treino foi de 86,6% e da matriz de teste de 88,3%.

| Treino | | | |
|---|---|---|---|
| *Regimes* | *1* | *2* | *3* |
| *1* | 792 | 91 | 11 |
| *2* | 79 | 697 | 43 |
| *3* | 59 | 58 | 728 |

| Teste | | | |
|---|---|---|---|
| *Regimes* | *1* | *2* | *3* |
| *1* | 304 | 28 | 9 |
| *2* | 34 | 345 | 18 |
| *3* | 19 | 20 | 320 |

TABELA 4.5 – Matrizes de Confusão de treino e teste do modelo CART

O modelo CART apresentou bom desempenho, mostrou-se fortemente influenciável pela variável temporal regime anterior, a qual está diretamente relacionada aos nós terminais 3 e 11, que representam 1410 de 2558 observações, ou seja, 55%.

#### 4.3.1.1 Florestas Aleatórias

O modelo de florestas aleatórias (*random forest* - RF), foi implementado utilizando-se o pacote *randomForest*. O valor do parâmetro *mtry* para este modelo que possui oito variáveis é $\sqrt{8}$, logo o valor padrão (2) é satisfatório, pois é o maior número inteiro cujo quadrado é menor ou igual a 8. Entretanto, como apontado no capítulo três, *mtry* é um dos principais parâmetros a ser otimizado. A Figura 4.21 apresenta o gráfico gerado pela otimização realizada, no qual o valor do erro OOB é menor para *mtry* igual a 2, confirmando o valor ideal para este parâmetro.

A Tabela 4.6 mostra as matrizes de treino e teste obtidas pelo modelo. Os erros percentuais, por classe, apresentados pelas matrizes de treino e teste são de 10,9% e 8,9% para o regime 1, 14,21% e 15,4% para o regime 2, 7,8% e 6,4% para o regime 3. A acurácia da matriz de treino é de 88,9% e da matriz de teste 89,7%.

| Treino | | | |
|---|---|---|---|
| *Regimes* | *1* | *2* | *3* |
| *1* | 769 | 86 | 9 |
| *2* | 78 | 730 | 43 |
| *3* | 26 | 40 | 777 |

| Teste | | | |
|---|---|---|---|
| *Regimes* | *1* | *2* | *3* |
| *1* | 329 | 26 | 6 |
| *2* | 39 | 320 | 17 |
| *3* | 3 | 20 | 338 |

TABELA 4.6 – Matrizes de Confusão de treino e teste do modelo RF

A Figura 4.22 apresenta a importância das variáveis no modelo RF. Apesar de RF não oferecer interpretação, observa-se que a variável regime anterior é a de maior importância, o que está em concordância com CART. Foi o modelo que apresentou melhor acurácia (quase 90%), se comparado a CART.

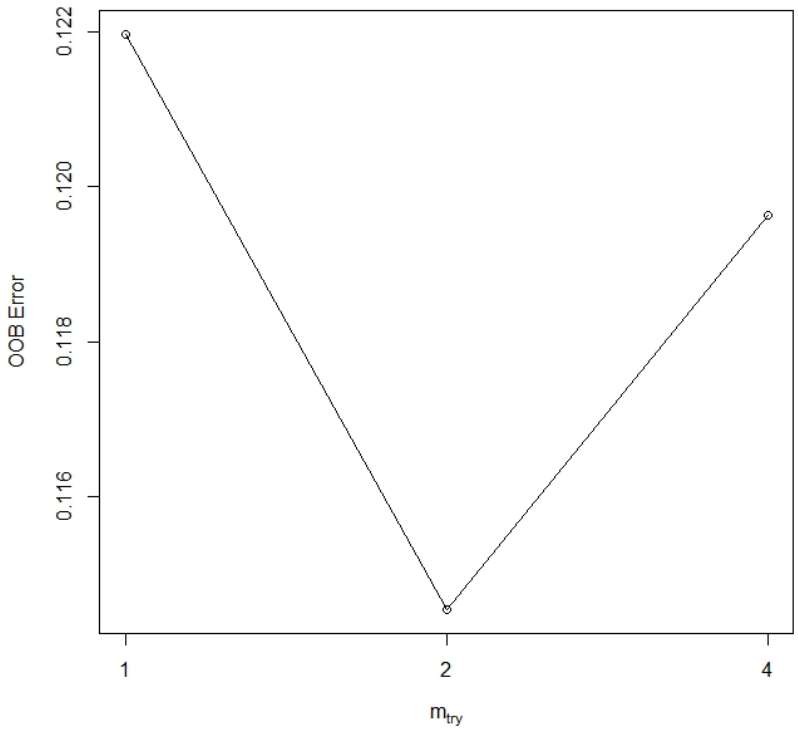



FIGURA 4.21 – Valor do parâmetro *mtry*

#### 4.3.1.2 Máquinas de Vetores de Suporte

Os parâmetros *custo* e *gamma* do modelo de Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM), gerados com os padrões pré-definidos do pacote *e1071*, valem 1 e 0,0384. A otimização do parâmetro *custo*, foi feita com testes para um primeiro intervalo de 1 a 2, com a variação de $2^{(} - 1 : 1)$, e um segundo intervalo de 1 a 3. Os valores de *gamma* foram testados simultaneamente no intervalo de 0,01 a 0,08 com a variação de 0,0025. Os valores que definiram os melhores parâmetros para o *custo* e *gamma* foram 2 e 0,025 para o primeiro intervalo e para o segundo intervalo 3 e 0,03. O erro para o primeiro intervalo foi de 0,1294 e para o segundo foi de 0,129. Diante de uma melhora muito pequena no desempenho do modelo, manteve-se o modelo com *custo* 2 e *gamma* 0,025.

As matrizes de confusão de treino e teste do modelo SVM são exibidas na Tabela 4.7. Os erros percentuais, por classe, apresentados mutuamente pelas matrizes de treino e teste são de 13,2% e 11,3% para o regime 1, 13% e 16,2% para o regime 2, 9,7% e 9,3% para o regimes 3. A acurácia da matriz de treino é de 88% e da matriz de teste 87,7%.

O modelo SVM não possibilita interpretação e apresentou acurácia de 87,7%, próxima do modelo RF. Observa-se que em todos os modelos, o regime 2 apresenta a maior taxa de

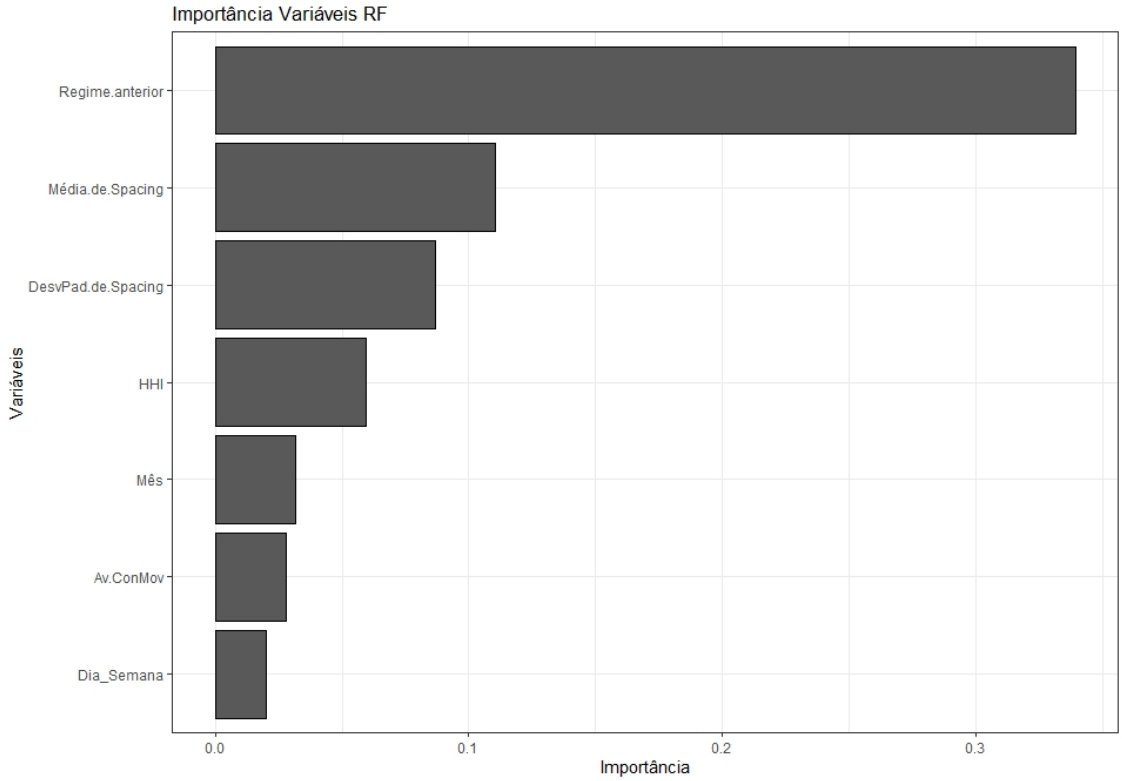



FIGURA 4.22 – Importância das variáveis no modelo RF

erro na matriz de confusão. E de acordo com o modelo CART, este é o único regime que sofre influência de outras variáveis independentes e apresenta alta probabilidade transição para outros regimes.

| Treino | | | |
|---|---|---|---|
| *Regimes* | *1* | *2* | *3* |
| *1* | 763 | 115 | 1 |
| *2* | 73 | 731 | 36 |
| *3* | 19 | 62 | 757 |

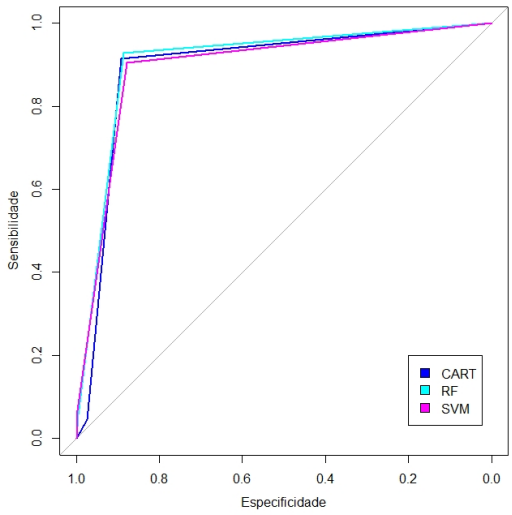
| Teste | | | |
|---|---|---|---|
| *Regimes* | *1* | *2* | *3* |
| *1* | 315 | 40 | 0 |
| *2* | 36 | 315 | 25 |
| *3* | 13 | 21 | 332 |

TABELA 4.7 – Matrizes de Confusão de treino e teste do modelo SVM

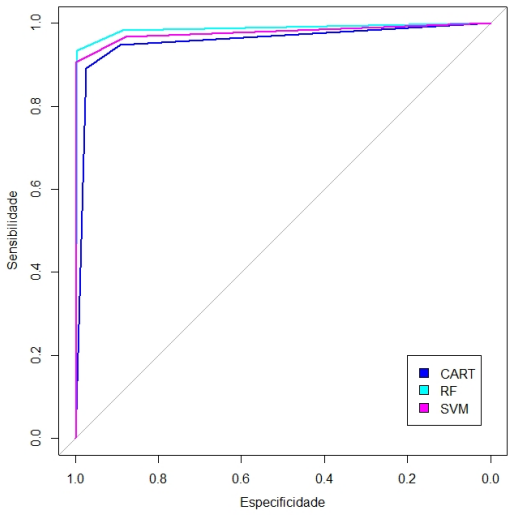
#### 4.3.1.3 Análise ROC

Para a análise ROC multiclasse foram gerados gráficos para a visualização do desempenho de cada classificador para cada regime. A Figura 4.23 mostra os três gráficos gerados. De modo geral, o desempenho dos classificadores para cada classe é similar, com uma leve superioridade do modelo RF. De acordo com a AUC, o RF foi o classificador com melhor desempenho, 0,942, seguido pelo SVM, 0,927 e CART, 0,917. Este resultado está

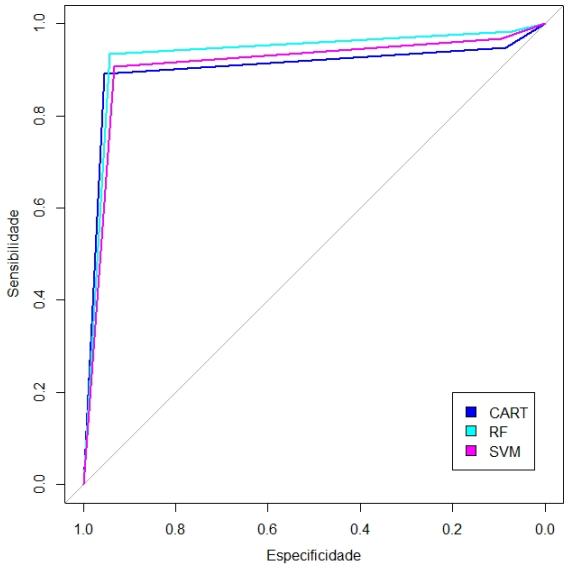
consoante com a acurácia de cada modelo, na qual RF obteve 89,7%, em seguida tem-se o SVM com 87,7% e CART 86,6%. Infere-se que os classificadores geraram modelos capazes de fazer previsões com boas margens de acerto.




(a) Regime 1 vs Regime 2




(b) Regime 1 vs Regime 3




(c) Regime 2 vs Regime 3

FIGURA 4.23 – Desempenho da curva ROC dos classificadores em cada regime

De acordo com os resultados obtidos, infere-se que a hipótese de que o estado atual depende do seu estado anterior se aplica a este problema, pois é um modelo fortemente influenciado pela variável temporal regime anterior. Assim, diante dos bons resultados, assume-se que o modelo é capaz de antecipar se o dia pertencerá aos regimes 1, 2 ou 3 com um dia de antecedência.

# 5 CONCLUSÃO

Neste Capítulo é feito um resumo desta pesquisa, identificado os principais métodos e suas implicações. O objetivo deste estudo foi a antecipação de dias congestionados no Aeroporto Internacional de Guarulhos. Para isto, em um primeiro momento, foi gerado um modelo por meio do método HMM. Em um segundo momento foram criados classificadores que geraram o modelo de previsão.

Utilizando a série temporal fatia diária de voos atrasados e cancelados, foram identificados três regimes a partir do modelo de HMM, definidos de acordo com a taxa média da fatia diária como: (i) muito congestionado; (ii) com congestionamento médio; e (iii) pouco congestionado. O critério de seleção do modelo com três regimes considerou os critérios AIC, BIC e LL. Os regimes identificados pelo HMM foram considerados como variável resposta do modelo de previsão gerado a partir dos classificadores.

O modelo de previsão foi gerado por meio de três classificadores CART, RF e SVM. CART foi escolhido por proporcionar a interpretação do modelo de previsão, do qual é possível identificar quais foram as variáveis selecionadas e sua significância para o modelo. RF foi escolhido porque fornece uma boa acurácia, é relativamente robusto a ruídos e fornece estimativas da importância das variáveis. O método SVM foi adotado por ser um método simples e eficiente de classificação, além da possibilidade da aplicação à base de dados sem a necessidade de redução dimensional.

O modelo HMM mostrou-se consistente ao observar os regimes estimados comparados à probabilidade posterior dos regimes. De modo geral foram os regimes foram detectados pela maior probabilidade de ocorrência de um determinado regime. Entretanto, o pacote utilizado faz o uso de probabilidades denominadas pelos autores como *smoothed probabilities*, assim, há situações em que a maior probabilidade é desconsiderada ao detectar um regime.

Os resultados obtidos pelos classificadores RF e SVM se mostraram coerentes com aqueles do CART. No CART foram identificadas as variáveis determinantes para dias muito congestionados e como estão combinadas. A árvore gerada possui seis nós terminais e é composta pelas variáveis regime anterior (gerada a partir da variável regime), média de *spacing*, dia da semana e HHI. A relevância das variáveis de RF estava em concordância

com a relevância das variáveis do CART. O modelo SVM não disponibiliza interpretação, entretanto a matriz de confusão gerada pelo modelo apresentou resultados similares aos modelos do CART e RF.

As curvas ROC geradas para problemas multiclasse e a os valores da AUC de cada modelo próximos a 1, corroboraram para confirmar que os modelos gerados pelos classificadores apresentam resultados satisfatórios. Os métodos de classificação se mostraram adequados e com boa acurácia para fazer a previsão de dias congestionados no Aeroporto Internacional de Guarulhos com um dia de antecedência. Os resultados obtidos são consistentes com a literatura e confirmou-se a hipótese da dependência temporal do modelo.

Algumas limitações foram identificadas no decorrer deste estudo:

- Os meses de junho e julho do ano de 2014 foram desconsiderados nesta análise, pois os dados não foram fornecidos pela ANAC;
- Ao detectar os regimes, considerando as *smoothed probabilities*, as maiores probabilidades que foram desconsideradas não foram analisadas;
- As variáveis obtidas a partir do resultado do modelo de HMM foram tidas como dados reais, assim, não foi considerada a incerteza dos rótulos;
- A variável Regime como premissa da variável Regime anterior pode levar os classificadores a considerá-la com maior peso em relação as demais variáveis;
- O modelo de previsão gerado, faz a antecipação da ocorrência de dias congestionados somente para o instante de tempo $t + 1$.

Para trabalhos futuros, propõe-se: (i)Pensar em alternativas à variável regime anterior; (ii) testar como variável resposta a probabilidade posterior dos regimes, levando em consideração a incerteza dos rótulos; (iii) investigar como outras variáveis independentes influenciariam o modelo de previsão de dias congestionados, pois a interpretação da combinação destes indicadores possibilita apontar possíveis alternativas para diminuir a taxa de atrasos e cancelamentos de voos; (vi) investigar a possibilidade de gerar previsões em um horizonte maior de tempo.

# Referências


ABDEL-ATY, M.; LEE, C.; BAI, Y.; LI, X.; MICHALAK, M. Detecting periodic patterns of arrival delay. **Journal of Air Transport Management**, v. 13, n. 6, p. 355–361, Nov 2007. 15, 16, 18, 20, 53

AMINIKHANGHAHI, S.; COOK, D. J. A survey of methods for time series change point detection. **Knowledge and Information Systems**, v. 51, n. 2, p. 339–367, May 2017. 22, 23, 24, 25

APREM, A.; KRISHNAMURTHY, V. Utility Change Point Detection in Online Social Media: A Revealed Preference Framework. **IEEE Transactions on Signal Processing**, v. 65, n. 7, p. 1869–1880, Apr 2017. 25

ASKARI, S.; MONTAZERIN, N.; ZARANDI, M. H. F. High-Frequency Modeling of Natural Gas Networks From Low-Frequency Nodal Meter Readings Using Time-Series Disaggregation. **IEEE Transactions on Industrial Informatics**, v. 12, n. 1, p. 136–147, Feb 2016. 22

ATER, I. Internalization of congestion at US hub airports. **Journal of Urban Economics**, v. 72, n. 2-3, p. 196–209, Sept 2012. 74

BAIK, H.; LI, T.; CHINTAPUDI, N. K. Estimation of Flight Delay Costs for U.S. Domestic Air Passengers. **Transportation Research Record:** Journal of the Transportation Research Board, v. 2177, n. 1, p. 49–59, 2010. 17

BALAKRISHNA, P.; GANESAN, R.; SHERRY, L. Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times : A case-study of Tampa Bay departures. **Transportation Research Part C:** Emerging Technologies, v. 18, n. 6, p. 950–962, 2010. 20, 21

BELOBABA, P.; ODONI, A.; BARNHART, C. **The Global Airline Industry**. 1. ed. United Kingdom: WILEY, 2009. 16

BENDINELLI, W. E.; BETTINI, H. F.; OLIVEIRA, A. V. Airline delays, congestion internalization and non-price spillover effects of low cost carrier entry. **Transportation Research Part A:** Policy and Practice, v. 85, p. 39–52, Mar 2016. 15, 16, 17

BRASIL. Departamento de Controle do Espaço Aéreo (DECEA). **Anuário estatístico de tráfego aéreo**. [S.l.], 2017. 18, 44, 46

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, Oct 2001. 36

BRODSKY, B.; DARKHOVSKY, B. **Mathematics and Its Applications**: Nonparametric methods in change point problems. 1. ed. Boston: Kluwer Academic Publishers, 1993. 25, 26, 27

BUXI, G.; HANSEN, M. Generating day-of-operation probabilistic capacity scenarios from weather forecasts. **Transportation Research Part C:** Emerging Technologies, v. 33, p. 153–166, Aug 2013. 21, 22

CHANDRAMOULEESWARAN, K. R.; KRZEMIEN, D.; BURNS, K.; TRAN, H. T. Machine Learning Prediction of Airport Delays in the US Air Transportation Network. In: AVIATION TECHNOLOGY, INTEGRATION AND OPERATIONS CONFERENCE, 2018. **Proceedings[...]**. [S.l.]: Reston:AIAA, 2018. p. 1–10. 21

CHEN, J.; GUPTA, A. K. **Parametric statistical change point analysis**: with applications to genetics, medicine, and finance. 2. ed. Boston: Birkhäuser Boston, 2012. 23, 25

COSTA, T. F.; LOHMANN, G.; OLIVEIRA, A. V. A model to identify airport hubs and their importance to tourism in Brazil. **Research in Transportation Economics**, v. 26, n. 1, p. 3–11, Jan 2010. 18

ETZ, A. Introduction to the Concept of Likelihood and Its Applications. **Advances in Methods and Practices in Psychological Science**, v. 1, n. 1, p. 60–69, Mar 2018. 31

EUROCONTROL. **Performance Review Report 2007:** an Assessment of Air Traffic Management in Europe during the Calendar Year 2007 : Eurocontrol performance review commission. [S.l.], 2008. 17

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: INTERNATIONAL CONFERENCE OF KNOWLEDGE DISCOVERY AND DATA MINING, 2.,1996. **Proceedings[...]**. [S.l.]: AAAI Press, 1996. p. 82–88. ix, 38, 39, 47, 50

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27–34, Nov 1996. 38

FERGUSON, J.; KARA, A. Q.; HOFFMAN, K.; SHERRY, L. Estimating domestic US airline cost of delay based on European model. **Transportation Research Part C:** Emerging Technologies, v. 33, p. 311–323, Aug 2013. 15, 17, 53

GALLAGHER, C.; LUND, R.; ROBBINS, M. Changepoint detection in climate time series with long-term trends. **Journal of Climate**, v. 26, n. 14, p. 4994–5006, Jul 2013. 22

GHASSEMPOUR, S.; GIROSI, F.; MAEDER, A. Clustering multivariate time series using hidden markov models. **International Journal of Environmental Research and Public Health**, v. 11, n. 3, p. 2741–2763, Mar 2014. 22, 27

GIRSHICK, M. A.; RUBIN, H. A Bayes approach to a quality control model. **The Annals of Mathematical Statistics**, v. 23, n. 1, p. 114–125, Mar 1952. 25

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. 2. ed. New York, NY: Springer New York, 2009. (Springer Series in Statistics). 35

INTERNATIONAL AIR TRANSPORT ASSOCIATION. **Diretrizes mundiais para slots**. Montreal: IATA, 2017. 74

INTERNATIONAL AIR TRANSPORT ASSOCIATION. **Annual Review 2019**. Montreal: IATA, 2019. 17

JACQUILLAT, A.; ODONI, A. R. An Integrated Scheduling and Operations Approach to Airport Congestion Mitigation. **Operations Research**, v. 63, n. 6, p. 1390–1410, Dec 2015. 15, 16

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. 1. ed. New York, NY: Springer, 2013. v. 103. (Springer Texts in Statistics, v. 103). 35, 37, 56

JAMES, N. A.; MATTESON, D. S. ECP : an R Package for nonparametric multiple change point analysis of multivariate data. **Journal of Statistical Software**, v. 62, n. 7, p. 334–345, Jan 2014. 24

JANIĆ, M. Modelling the resilience, friability and costs of an air transport network affected by a large-scale disruptive event. **Transportation Research Part A: Policy and Practice**, v. 71, p. 1–16, Jan 2015. 16

KANDHASAMY, J. P.; BALAMURALI, S. Performance analysis of classifier models to predict diabetes mellitus. **Procedia Computer Science**, v. 47, p. 45–51, 2015. 36

KARATZOGLOU, A.; MEYER, D.; HORNIK, K. Support vector machines in R. **Journal of Statistical Software**, v. 15, n. 9, p. 1–28, Apr 2006. 37

KILLIC, R.; ECKLEY, A. I. changepoint : An r package for changepoint analysis. **Journal of Statistical Software**, v. 58, n. 3, p. 1–19, Jun 2014. 24

KUHA, J. AIC and BIC:comparisons of assumptions and performance. **Sociological Methods & Research**, v. 33, n. 2, p. 188–229, Nov 2004. 34, 35

LIU, P.-c. B.; HANSEN, M.; MUKHERJEE, A. Scenario-based air traffic flow management: From theory to practice. **Transportation Research Part B: Methodological**, v. 42, n. 7-8, p. 685–702, Aug 2008. 21

LORENA, A. C.; De Carvalho, A. C. P. L. F. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, Dec 2007. 37

LOURENCO, L. F. N.; SALLES, M. B. de C.; GEMIGNANI, M. M. F.; GOUVEA, M. R.; KAGAN, N. Time Series modelling for solar irradiance estimation in northeast Brazil. In: INTERNATIONAL CONFERENCE ON RENEWABLE ENERGY RESEARCH AND APPLICATIONS, 6., 2017. **Proceedings [...]**. [S.l.]: Piscataway: IEEE, 2017. p. 401–405. 22

LUONG, T. M.; PERDUCA, V.; NUEL, G. **Hidden Markov model applications in change-point analysis**. [Ithaca, NY: Computer Society, Cornell University], 2012. Disponível em: <https://arxiv.org/pdf/1212.1778.pdfAcessoem:15out.2019.> 27

LYSTIG, T. C.; HUGHES, J. P. Exact computation of the observed information matrix for Hidden Markov Models. **Journal of Computational and Graphical Statistics**, v. 11, n. 3, p. 678–689, Sept 2002. 32

MADAS, M. A.; ZOGRAFOS, K. G. Airport capacity vs . demand : Mismatch or mismanagement ? **Elsevier Transportation Research Part A:** Policy and Practice, v. 42, p. 203–226, 2008. 16

MAIMON, O.; ROKACH, L. Introduction to knowledge discovery and data mining. In: MAIMON, O.; ROKACH, L. (ed.). **Data mining and knowledge discovery handbook**. 2nd. ed. [S.l.]: New York: Springer, 2010. p. 1–15. 38

MAINDONALD, J.; BRAUN, W. J. **Data analysis and graphics using R**: [ An Example-Based Approach]. 3. ed. New York: Cambridge University Press, 2010. 55

MEI, Y. Sequential change-point detection when unknown parameters are present in the pre-change distribution. **The Annals of Statistics**, v. 34, n. 1, p. 92–122, Feb 2006. 25

MISHIN, D.; BRANTNER-MAGEE, K.; CZAKO, F.; SZALAY, A. S. Real time change point detection by incremental PCA in large scale sensor data. In: IEEE HIGH PERFORMANCE EXTREME COMPUTING CONFERENCE, 2014. **Proceedings[...]**. [S.l.]: Piscataway:IEEE, 2014. p. 1–6. 25

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de Séries Temporais**. 2. ed. São Paulo: Editora Blucher, 2006. 22

NEXTOR; BALL, M.; BARNHART, C.; DRESNER, M.; NEELS, K.; ODONI, A.; PETERSON, E.; SHERRY, L.; TRANI, A.; ZOU, B. **Total delay impact study a comprehensive assessment of the costs and impacts of flight delay in the United States**. Berkeley, 2010. 82 p. (Technical reports). 17

PAGE, E. S. Continuous Inspection Schemes. **Biometrika**, v. 41, n. 1-2, p. 100–115, Jun 1954. 25

PYRGIOTIS, N.; MALONE, K. M.; ODONI, A. Modelling delay propagation within an airport network. **Transportation Research Part C:**Emerging Technologies, v. 27, p. 60–75, 2013. 17

REBOLLO, J. J.; BALAKRISHNAN, H. Characterization and prediction of air traffic delays. **Transportation Research Part C:** Emerging Technologies, v. 44, p. 231–241, 2014. 17, 18, 21, 45, 53

SALMON, M.; SCHUMACHER, D.; HÖHLE, M. Monitoring count time series in R : aberration detection in public health surveillance. **Journal of Statistical Software**, v. 70, n. 10, p. 1–35, May 2016. 22

SANTOS, G.; ROBIN, M. Determinants of delays at European airports. **Transportation Research Part B:** Methodological, v. 44, n. 3, p. 392–403, Mar 2010. 16, 53

SANTOS, T. A. dos; VENDRAME, I.; ALVES, C. J. P.; CAETANO, M.; SILVA, J. P. S. Modelo de identificação do impacto futuro de chuvas extremas nos atrasos/cancelamentos de voos. **TRANSPORTES**, v. 26, n. 2, p. 44–53, Aug 2018. 16

SCARPEL, R. A. A demand trend change early warning forecast model for the city of São Paulo multi-airport system. **Transportation Research Part A:** Policy and Practice, v. 65, p. 23–32, 2014. 20, 21, 35, 36

SCARPEL, R. A.; PELICIONI, L. C. A data analytics approach for anticipating congested days at the São Paulo International Airport. **Journal of Air Transport Management**, v. 72, p. 1–10, Sept 2018. 16, 18, 19, 20, 21, 53, 74

SHEWHART, W. A. Quality control charts. **Bell System Technical Journal**, n. 5, p. 593–603, Oct 1926. 24

SHUMWAY, R. H.; STOFFER, D. S. **Time series analysis and its applications**. 4. ed. Cham: Springer, 2017. (Springer Texts in Statistics). 22

STERNBERG, A.; SOARES, J.; CARVALHO, D.; OGASAWARA, E. **A review on flight delay prediction**. [Ithaca, NY: Computer Society, Cornell University], v. 2, p. 1–21, Mar 2017. Disponível em: <https://arxiv.org/pdf/1703.06118.pdfAcessoem:15out.2019.> 20

TARTAKOVSKY, A.; NIKIFOROV, I.; BASSEVILLE, M. **Sequential analysis**: Hypothesis testing and changepoint detection. 1. ed. New York: CRC Press, 2015. 25

TZOURAS, S.; ANAGNOSTOPOULOS, C.; MCCOY, E. Financial time series modeling using the Hurst exponent. **Physica A:** Statistical Mechanics and its Applications, v. 425, p. 50–68, May 2015. 22

UNITED STATES. Congress. Joint Economic Committee. **Your flight has been delayed again**: flight delays cost passengers, airlines, and the U.S. economy billions. Washington, DC: Joint Economic Committee, Senate, 2008. 17

VISSER, I. Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series. **Journal of Mathematical Psychology**, v. 55, n. 6, p. 403–415, Dec 2011. 29, 30, 31

VISSER, I.; RAIJMAKERS, M.; MAAS, H. Hidden markov models for individual time series. In: VALSINER J.; MOLENAAR P.; LYRA M.; CHAUDHARY N. (ed). **Dynamic process methodology in the social and developmental sciences**. [S.l.]: New York: Springer, 2009. Chap. 13. p. 269–289. 50

VISSER, I.; SPEEKENBRINK, M. depmixS4 : An R Package for Hidden Markov Models. **Journal of Statistical Software**, v. 36, n. 7, p. 1–21, Aug 2010. 29, 32, 33, 34

WARREN, L. T. Clustering of time series data survey. **Pattern Recognition**, v. 38, n. 11, p. 1857–1874, Nov 2005. 22, 23

WEI, W.; LIU, C.; ZHU, M. Y.; MATEI, S. A. A Non-parametric Hidden Markov clustering model with applications to time varying user activity analysis. In: IEEE INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATION, 14., 2015, Miami. **Proceedings [...]**. [S.l.]: Piscataway: IEEE, 2015. p. 549–554. 22, 27

WENSVEEN, J. G. **Air Transportation**: a management perspective . 8. ed. New York, NY: Routledge, 2016. 17, 18

WOOLDRIDGE, J. M. **Introductory Econometrics**: a modern approach. 6. ed. Boston: Cengage Learning, 2016. 23

XIONG, J.; HANSEN, M. Modelling airline flight cancellation decisions. **Transportation Research Part E:** Logistics and Transportation Review, v. 56, p. 64–80, Sept 2013. 15, 20

YU, B.; GUO, Z.; ASIAN, S.; WANG, H.; CHEN, G. Flight delay prediction for commercial air transport : A deep learning approach. **Transportation Research Part E:** Logistics and Transportation Review, v. 125, n. March, p. 203–221, 2019. 20, 21

ZUCCHINI, W.; MACDONALD, I. L.; LANGROCK, R. **Hidden Markov models for time series**: an introduction using r, second edition. 2. ed. London: Chapman and Hall/CRC, 2016. ix, 22, 28, 29, 30, 31, 32, 33, 34, 50

# Apêndice A - Código

## A.1 Apêndice A

# Código

**Os códigos utilizados no decorrer deste trabalho, estão disponibilizados de forma generalizada.**

## Pré-processamento

### Conjunto de Dados

**Com o propósito de facilitar o pré-processamento, as variáveis explicativas HHI, média.de.Spacing, DesvPad.de.Spacing e Av.ConMov já foram inclusas.**

**Alterando o formato da coluna data**

data.gru1 $ Data.1 <- as.Date(data.gru1 $ Data.1, format = "%d/%m/%Y")

names(data.gru1)[1] <-"Data" #renomeando coluna

### Valores Faltantes

**Tratando valores faltantes**

library(mice)

md.pattern(data.gru) #identificar valores faltantes

library(VIM)

missingvalue <- data.gru[,-c(1,6,7,8,9)]

names(missingvalue)[3] <- "Atrasado" #renomeando coluna

matrixplot(missingvalue) #plot de dados faltantes

is.na(data.gru) #observação faltante: 1037 de "Cancelado"

data.gru1 <- na.omit(data.gru) Omitir dados não numéricos (só há 1)

na.fail(data.gru1) # verificando se o valor faltante foi omitido

### Análise Prévia dos Dados

**Gráfico dos voos realizados e atrasados**

data.delay <- data.frame(data.gru1$Data, data.gru1$Realizado) #dataframe

colnames(data.delay) <- c("Data", " Realizado") #renomeando colunas

data.delay ["Atrasado"] <- data.frame(data.gru1$Realizado.com.atraso + data.gru1$Cancelado) #adicionando coluna "Atrasado"

**Soma de cada coluna**

Realizado <- sum(data.delay$Realizado)

Atrasado <- sum(data.delay$Atrasado)

Total <- sum(data.gru1$Movimentos.Programados)

**Criando dataframe com o total das variáveis Realizados e Cancelados**

library(reshape2) #função melt

data.delay<- data.frame("Sem.atraso" = Realizado, "Atrasado.cancelado" = Atrasado)

data.delay.melt <- melt(data.delay)

**Adcionando porcentagens ao dataset melted**

library(dplyr) #funçaõ mutate

data.delay.group<-mutate(group_by(data.delay.melt, variable, percent = (value / sum(value))) ) %>% ungroup()

**Plot**

library(ggplot2) #plot

library(scales) # escalas de porcentagem

p0 <-ggplot(data.delay.group, aes(x=variable, y=value, fill= factor(variable, labels = c("Sem atraso", "Atrasado/Cancelado"))))+

geom_bar(stat = "identity", width = .5)+ # plotar x e y

labs(x= "Voos", y="Frequência")+

ggtitle("Movimentos de Voos")+ theme_minimal()+

scale_fill_grey()+ #cor do gráfico desejada

geom_text(aes(label = percent(percent)), vjust=-0.3 , size=4)+ #localização labels

theme(text= element_text(size=15),legend.title = element_blank(), axis.title.x=element_blank())+ #retirar título da legenda

scale_x_discrete(labels= (c("Sem.atraso" = "Sem atraso", "Atrasado.cancelado" = "Atrasado/Cancelado")))

**Foram ainda gerados gráficos anual, mensal, semanal, dias úteis e final de semana.**

**Para gerar as variáveis de cada gráfico, foram utilizados os seguintes pacotes:**

library(data.table) #função serDT, extrai as variáveis ano, mês e semana do conjunto de dados.

library(reshape2) #função melt, faz a transposição do conjunto de dados

library(reshape) #função cast, soma as linhas das colunas do conjunto de dados

## Identificando e removendo outliers

**Renomeando coluna Total.geral**

colnames(data.gru1)[colnames(data.gru1)=="Total.geral"] <- "Movimentos.Programados"

data.gru.vars <- data.gru1[,c(6,7,8,9)] #Conjunto de dados contendo as variáveis explicativas.

**Ordenando colunas da base que vai gerar a fatia (porcentagem) de voos atrasados e canceldos sem as variáveis explicativas.**

data.gru1 = data.gru1[, c(1,3,4,2,5)]

data.fails<- data.frame(data.gru1$Realizado.com.atraso + data.gru1$Cancelado) / data.gru1$Movimentos.Programados

colnames(data.fails)= c("Fatia") #nomeando coluna

**Adicionando coluna data**

Data <- data.gru1[,1] #variável Data

data.fails["Data"]<-c(Data)#adcionando Data ao conjunto de dados

data.fails = data.fails[,c(2,1)] #reordenando

data.fails.vars <- cbind(data.fails, data.gru.vars)

**Boxplot com outlier**

plo.box.gg.vars <- data.frame(data.fails.vars[,-1])

out<-ggplot(plo.box.gg.vars, aes("", Fatia))+

geom_boxplot(fill="gray")+

labs(x= "")+

theme_bw()

**Identificando e removendo outliers**

outliers <- boxplot(data.fails.vars$Fatia, plot=F)$out #variável contento outliers, 85 outliers

data.fails.vars[which(data.fails.vars$Fatia %in% outliers),] #Identificando outliers

data.fails.vars <- data.fails.vars[-which(data.fails.vars$Fatia %in% outliers),]#removendo outliers

**Boxplot sem outlier**

nout.box.gg.vars <- data.frame(data.fails.vars[,-1])#variável para plotar

nout<-ggplot(nout.box.gg.vars, aes("", Fatia))+

geom_boxplot(fill="gray")+

labs(x= "")+

theme_bw()

**Plot**

library(ggpubr)

ggarrange(out, nout, labels = c("A", "B"))

# Modelos

## HMM

set.seed(7)

library(depmixS4)

data.fail = data.frame(data.fails[,-1])

colnames(data.fail)= c("Fatia") #nomeando coluna

hmm.model3 <- depmix(Fatia~1, family = gaussian(), nstates = 3, data=data.fail) #modelo

hmm.fit3 <- fit(hmm.model3, emcontrol=em.control(maxit = 5000)) #Foram ainda gerados modelos com 2 regimas a 6 regimes.

### Gráfico AIC e BIC, número de regimes.

x11(width=10)

par(mfrow=c(1,2))

plot(2:6, c(AIC(hmm.fit2),AIC(hmm.fit3), AIC(hmm.fit4),AIC(hmm.fit5), AIC(hmm.fit6)), ty = "b", ylab = "AIC", xlab = "Número de grupos", main= "AIC", col="gray", lwd=2, panel.first = grid(col = "gray"))

plot(2:6, c(BIC(hmm.fit2),BIC(hmm.fit3), BIC(hmm.fit4),BIC(hmm.fit5), BIC(hmm.fit6)), ty = "b", ylab = "BIC", xlab = "Número de grupos", main = "BIC", col="gray", lwd=2, panel.first = grid(col = "gray"))

### Gráfico probabilidade posterior

post.prob3 <- posterior(hmm.fit3) # Probabilidades posteriores e regimes

X11(width = 10)

layout(1:2)

matplot(post.prob3[,-1], type = "l", ylab = "Probabilidade", xlab= "Dias", main="", col = c("red", "yellow", "green"))

legend(x='bottomright', c('R1','R2', 'R3'), fill=c("red", "yellow", "green"), bty='n', bg= "gray90", box.col = "green4", xjust=0)

hmmregime <- post.prob3[,1] #Regimes estimados

### Plot série temporal fatia, regimes e probabilidade posterior dos primeiros 250 dias de 2011.

X11(width = 10)

layout(1:3)

plot(data.fails[1:250,2], type = "l", ylab="Fatia", xlab="", col="gray", panel.first = grid(col = "gray"), main="Série Temporal Fatia")

plot(post.prob3$state[1:250], type='h', main='Regimes Estimados', xlab='', ylab='Regimes', col=c("red", "yellow", "green")[(hmmregime[1:338,])], lwd=2)

legend(x='topleft', c('R1','R2', 'R3'), fill=c("red", "yellow", "green"), bty='n', xjust=0)

matplot(post.prob3[1:250,-1], type = "l", ylab = "Probabilidade", xlab= "Dias (2011)", main = "Probabilidade Posterior dos Regimes", col = c("red", "yellow", "green"))

legend(x='topleft', c('R1','R2', 'R3'), fill=c("red", "yellow", "green"), bty='n', xjust=0)

**Deste modo também foram gerados gráficos dos "Regimes Estimados"" e "Probabilidade Posterior dos Regimes" anualmente.**

## CART

**Foi utilizado o cojunto de dados da variável "data.fails.vars" , em que foram acrescidas as variáveis "Regime" (regimes detectados pelo HMM), "Regime.anterior" (variável obtida a partir da "Regime"), "Dia_semana" e "Mês" (extraídas da base de dados utilizando os comandos weekday e month do r). A variável "Fatia" foi retirada da base.**

summary(data.gru.class.vars) #conjunto de dados.

library(UBL)

data.gru.balanced<- AdasynClassif(Regime~.,data.gru.class.vars, dist= "HEOM") #Dados baleanceados.

**Amostras de treino e teste do modelo, 70% dos dados para treinamento e 30% para teste.**

set.seed(002) amostra <- sample.int(n = nrow(data.gru.balanced), size = floor(.7*nrow(data.gru.balanced)), replace=F)

treino <- data.gru.balanced[amostra,]

teste <- data.gru.balanced[-amostra,]

library(rpart)

fit <- rpart(Regime~., data=treino) #modelo

**Validação cruzada, 10-fold.**

printcp(fit)

**Poda da árvore**

bestcp <- fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"]

poda_fit <- prune(fit, cp=bestcp)

**Plot da árvore**

library(partykit)

plot(as.party(poda_fit), type="extended", use.n=F, ylab="Classificação")

## RF

set.seed(050)

**Criando bases de treino e teste**

library(caTools)

split.rf <- sample.split(data.gru.balanced$Regime, SplitRatio = 0.7)

treino.rf <- subset(data.gru.balanced, split.rf==TRUE)

teste.rf<- subset(data.gru.balanced, split.rf==F)

library(randomForest)

model.rf.tree <- randomForest(Regime~., data=treino.rf, importance=T, ntree=400) #modelo

**Avaliação dos resultados**

pred.test <- predict(model.rf.tree, teste.rf, type = “class”)

table(pred.test, teste.rf$Regime) #matriz de confusão

mean(pred.test == teste.rf$Regime) #acurácia

**Otimização do parâmetro mtry**

t.rf <- tuneRF(treino.rf[,-3], treino.rf[,3], plot=T, ntreeTry = 400, trace=T)

**Plot número de árvores**

windows()

plot(model.rf, main=“Modelo”, col=c(“blue”, “red”, “yellow”, “green”)) legend(“topright”, colnames(model.rf$err.rate), col=c(“blue”, “red”, “yellow”, “green”), cex=0.8, fill=c(“blue”, “red”, “yellow”, “green”))

**De acordo com o plot do número de árvores, foram treinados modelos com 400 e 500 árvores.**

## SVM

data.svm <-data.gru.class.vars[,c(-1)]

**Criando dummies**

library(dummies)

dummies <- dummy.data.frame(data.svm, names = c(“Dia_Semana”, “Mês”, “Regime.anterior”)) table(dummies$Regime)

**Balanceamento base de dados**

library(UBL)

dummies.balanced<- AdasynClassif(Regime~.,dummies, dist= “HEOM”)

table(dummies.balanced$Regime)

set.seed(005)

sample.svm <- sample.int(n = nrow(dummies.balanced), size =floor(.7*nrow(dummies.balanced)), replace=F)

treino.svm <- dummies.balanced[sample.svm,]

teste.svm <- dummies.balanced[-sample.svm,]

library(e1071)

svm.model <- svm(Regime~., data=treino.svm, cross=10, probability=T) #modelo

**Foi feita a avalização dos resultados gerando a matriz de confusão e verificando a acurácia.**

### Otimização dos parâmetros

svm.tune.result <- tune(svm, Regime~., data=treino.svm, kernel="radial",ranges = list(gamma=seq(0.01, 0.08, 0.0025), cost=2^(-1:1)))

## ROC CURVE

### Curva Roc multiclasse

```
library(pROC)

cart.pred.Teste <- predict(poda_fit, teste, type = "vector", ordered=T)

cart<-(multiclass.roc(teste$Regime, cart.pred.Teste, plot=F, col=4))

rf.pred.test <- predict(model.rf.tree, teste.rf, type = 'response', ordered=T)

rf <-multiclass.roc(teste.rf$Regime~as.numeric(rf.pred.test), col=5, plot=F)

svm.pred.test <-predict(svm.tune.result$best.model, teste.svm, ordered=T )

svm <-multiclass.roc(teste.svm$Regime~as.numeric(svm.Teste) , plot=F, col=6)

cart.rs <- cart["rocs"]]

rf.rs <- rf[["rocs"]]

svm.rs <- svm[["rocs"]]
```

### Plot da comparação das curvas

```
x11(width=20)

par(mfrow=c(1,3))

plot.roc(cart.rs[[1]], main="", col=5, xlab = "Especificidade", ylab="Sensibilidade", cex.lab=1.7)
plot.roc(rf.rs[[1]],add=T, col=6 )

plot.roc(svm.rs[[1]], add=T, col=7)

legend(0.2, 0.2, c('CART', 'RF', 'SVM'), 5:7, cex=1.5)

plot.roc(cart.rs[[2]], main="", col=5, xlab = "Especificidade", ylab="Sensibilidade", cex.lab=1.7)

plot.roc(rf.rs[[2]],add=T, col=6)

plot.roc(svm.rs[[2]], add=T, col=7)

legend(0.2, 0.2, c('CART', 'RF', 'SVM'), 5:7, cex=1.5)

plot.roc(cart.rs[[3]], main="", col=5, xlab = "Especificidade", ylab="Sensibilidade", cex.lab=1.7)

plot.roc(rf.rs[[3]],add=T, col=6)

plot.roc(svm.rs[[3]], add=T, col=7)

legend(0.2, 0.2, c('CART', 'RF', 'SVM'), 5:7, cex=1.5)
```

# Anexo A - Links para acesso às base de dados

## A.1 Anexo A

Link de acesso à base de dados Voo Regulare Ativo (VRA)) disponível no site da Agência Nacional de Aviação Civil (ANAC):

https://www.anac.gov.br/assuntos/dados-e-estatisticas/historico-de-voos

Link de acesso ao conjunto de dados utilizados no modelo de classificação:

https://drive.google.com/open?id=1ZbvswdZkRmNrPGh1TO5P-9yZVKveStxi

FOLHA DE REGISTRO DO DOCUMENTO

| 1. CLASSIFICAÇÃO/TIPO | 2. DATA | 3. DOCUMENTO Nº | 4. Nº DE PÁGINAS |
|---|---|---|---|
| DM | 18 de dezembro de 2019 | DCTA/ITA/DM-096/2019 | 97 |

5. TÍTULO E SUBTÍTULO:

ANTECIPAÇÃO DE MUDANÇA DE REGIME NA FATIA DIÁRIA DE VOOS ATRASADOS E CANCELADOS NO AEROPORTO INTERNACIONAL DE GUARULHOS

6. AUTORA(ES):

**Rosana Batista Teixeira**

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica – ITA

8. PALAVRAS-CHAVE SUGERIDAS PELA AUTORA:

Detecção de pontos de mudança; Hidden Markov models; Classificadores.

9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Detecção de transformação; Modelos escondidos de Markov; Classificações; Operações de linhas aéreas; Pesquisa operacional.

10. APRESENTAÇÃO: (**X**) **Nacional** ( ) **Internacional**

ITA, São José dos Campos. Curso de Mestrado. Programa de Pós-Graduação em Pesquisa Operacional. Área Métodos de Otimização. Orientador: Prof Dr. Rodrigo Arnaldo Scarpel. Defesa em 09/12/2019. Publicada em 2019.

11. RESUMO:

Atrasos e cancelamentos de voos são ocorrências frequentes na maioria dos aeroportos em todo o mundo. No Brasil, o aumento desregulamentado do tráfego aéreo provocou a concentração de voos em alguns aeroportos e possibilitou a ocorrência de atrasos e cancelamentos de voos em razão de dias congestionados. Dentre estes aeroportos, o Aeroporto Internacional de Guarulhos (GRU) é o mais afetado por atrasos no país. Portanto, o objetivo deste trabalho é a criação de um modelo de previsão que visa antecipar a ocorrência de dias congestionados no Aeroporto Internacional de Guarulhos. Para a composição do modelo foram empregues os Modelos Escondidos de Markov, como uma abordagem de agrupamento, e três classificadores: Árvore de Classificação e Regressão, Florestas Aleatórias e Máquina de Vetores de Suporte. A precisão do modelo foi considerada satisfatória e antecipou a mudança de regime na fatia diária por um período a frente.

12. GRAU DE SIGILO:

(**X**) **OSTENSIVO** ( ) **RESERVADO** ( ) **SECRETO**